Deployment and Scheduling of Wireless Sensor Networks for Air Pollution Monitoring

Ahmed Boubrima

N° d'ordre NNT :

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
L'INSA de Lyon

Ecole Doctorale ED512
InfoMaths

Spécialité/ discipline de doctorat :
Informatique

Soutenue publiquement le 12/03/2019, par :
Ahmed Boubrima

Deployment and Scheduling of Wireless Sensor Networks for Air Pollution Monitoring

---

Devant le jury composé de :

DIAS DE AMORIM, Marcelo	Directeur de recherche, CNRS	Rapporteur
DUDA, Andrzej	Professeur des universités, Grenoble INP-ENSIMAG	Rapporteur
CARNEIRO VIANA, Aline	Chargée de recherche HDR, INRIA	Examinatrice
GUERIN-LASSOUS, Isabelle	Professeur des universités, Univ. Lyon 1	Examinatrice
NOEL, Thomas	Professeur des universités, Université de Strasbourg	Examinateur
RIVANO, Hervé	Professeur des universités, INSA de Lyon	Directeur de thèse
BECHKIT, Walid	Maitre de conférences, INSA de Lyon	Co-directeur de thèse
CHAPPAZ, Claire	Atmo Auvergne-Rhône-Alpes	Invitée

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage **secretariat@edchimie-lyon.fr** INSA : R. GOURDON	M. Stéphane DANIELE Institut de recherches sur la catalyse et l'environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 Avenue Albert EINSTEIN 69 626 Villeurbanne CEDEX **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec. : M.C. HAVGOUDOUKIAN **ecole-doctorale.eea@ec-lyon.fr**	M. Gérard SCORLETTI École Centrale de Lyon 36 Avenue Guy DE COLLONGUE 69 134 Écully Tél : 04.72.18.60.97 Fax 04.78.43.37.17 **gerard.scorletti@ec-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : H. CHARLES **secretariat.e2m2@univ-lyon1.fr**	M. Philippe NORMAND UMR 5557 Lab. d'Ecologie Microbienne Université Claude Bernard Lyon 1 Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://www.ediss-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : M. LAGARDE **secretariat.ediss@univ-lyon1.fr**	Mme Emmanuelle CANET-SOULAS INSERM U1060, CarMeN lab, Univ. Lyon 1 Bâtiment IMBL 11 Avenue Jean CAPELLE INSA de Lyon 69 621 Villeurbanne Tél : 04.72.68.49.09 Fax : 04.72.68.49.16 **emmanuelle.canet@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 Fax : 04.72.43.16.87 **infomaths@univ-lyon1.fr**	M. Luca ZAMBONI Bât. Braconnier 43 Boulevard du 11 novembre 1918 69 622 Villeurbanne CEDEX Tél : 04.26.23.45.52 **zamboni@maths.univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Marion COMBE Tél : 04.72.43.71.70 Fax : 04.72.43.87.12 Bât. Direction **ed.materiaux@insa-lyon.fr**	M. Jean-Yves BUFFIÈRE INSA de Lyon MATEIS - Bât. Saint-Exupéry 7 Avenue Jean CAPELLE 69 621 Villeurbanne CEDEX Tél : 04.72.43.71.70 Fax : 04.72.43.85.28 **jean-yves.buffiere@insa-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Marion COMBE Tél : 04.72.43.71.70 Fax : 04.72.43.87.12 Bât. Direction **mega@insa-lyon.fr**	M. Jocelyn BONJOUR INSA de Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69 621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **http://ed483.univ-lyon2.fr** Sec. : Viviane POLSINELLI Brigitte DUBOIS INSA : J.Y. TOUSSAINT Tél : 04.78.69.72.76 **viviane.polsinelli@univ-lyon2.fr**	M. Christian MONTES Université Lyon 2 86 Rue Pasteur 69 365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Acknowledgements

This Ph.D. has been funded by an INRIA CORDI-S scholarship and supervised by Professors Walid Bechkit and Hervé Rivano within the Agora team of the CITI-lab of INSA-Lyon. It is also the core of the UrPolSens multidisciplinary project which is funded by the "Labex IMU" of the University of Lyon. The WSN scheduling part of this thesis has been funded in part by the Canadian Mitacs Globalink Research Award and took place at the University of Ottawa, Canada from Feb. to May 2018 as part of an international collaboration with Professor Azzedine Boukerche.

Abstract

Wireless sensor networks (WSN) are widely used in environmental applications where the aim is to sense a physical phenomenon such as temperature, humidity, air pollution, etc. In this context of application, the use of WSN allows to understand the variations of the phenomenon over the monitoring region and therefore be able to take adequate decisions regarding the impact of the phenomenon. Due to the limitations of its traditional costly monitoring methods in addition to its high spatial and temporal variability, air pollution is considered as one of the main physical phenomena that still need to be studied and characterized.

In this thesis, we consider three main applications regarding the use of WSN for air pollution monitoring: 1) the construction of real time air quality maps using sensor measurements; 2) the detection of pollution threshold crossings; and 3) the correction of physical models that simulate the pollution dispersion phenomenon. All these applications need careful deployment and scheduling of sensors in order to get a better knowledge of air pollution while ensuring a minimal deployment cost and a maximal lifetime of the deployed sensor network. Our aim is to tackle the problems of WSN deployment and scheduling while considering the specific characteristics of the air pollution phenomenon. We propose for each application case a new efficient approach for the deployment of sensor and sink nodes. We also propose a WSN scheduling approach that is adapted to the case of physical models' correction. Our optimization approaches take into account the physical nature of air pollution dispersion and incorporate real data provided by the existing pollution sensing platforms. As part of each approach, we use integer linear programming to derive optimization models that are well adapted to solving small and medium instances. To deal with large instances, we propose heuristic algorithms while using linear relaxation techniques.

Besides our theoretical works on air pollution monitoring, we design from scratch and deploy in the Lyon city a cost-effective energy-efficient air pollution sensor network. Based on the characteristics of our monitoring system in addition to real world air pollution datasets, we evaluate the effectiveness of our deployment and scheduling approaches and provide engineering insights for the design of WSN-based air pollution monitoring systems. Among our conclusions, we highlight the fact that the size of the optimal sensor network depends on the degree of the variations of pollution concentrations within the monitoring region.

Résumé

Les réseaux de capteurs sans fil (RCSF) sont largement utilisés dans les applications environnementales où l'objectif est de détecter un phénomène physique tel que la température, l'humidité, la pollution de l'air, etc. Dans ce contexte d'application, l'utilisation de RCSF permet de comprendre les variations du phénomène et donc être en mesure de prendre des décisions appropriées concernant son impact. En raison des limitations de ses méthodes de suivi traditionnelles et de sa grande variabilité spatiale et temporelle, la pollution de l'air est considérée comme l'un des principaux phénomènes physiques qui restent à étudier et à caractériser.

Dans cette thèse, nous considérons trois applications concernant l'utilisation de RCSF pour le suivi de la pollution de l'air : la cartographie en temps réel de la qualité de l'air, la détection de dépassements de seuils des polluants et la correction de modèles physiques qui simulent le phénomène de dispersion de la pollution. Toutes ces applications nécessitent de déployer et d'ordonnancer minutieusement les capteurs afin de mieux comprendre la pollution atmosphérique tout en garantissant un coût de déploiement minimal et en maximisant la durée de vie du réseau. Notre objectif est de résoudre les problèmes de déploiement et d'ordonnancement tout en tenant compte des caractéristiques spécifiques du phénomène de la pollution de l'air. Nous proposons pour chaque cas d'application une approche efficace pour le déploiement de noeuds capteurs et puits. Nous proposons également une approche d'ordonnancement adaptée au cas de la correction de modèles physiques. Nos approches d'optimisation prennent en compte la nature physique de la pollution atmosphérique et intègrent les données réelles fournies par les plateformes existantes de suivi de la qualité de l'air. Dans chacune de nos approches d'optimisation, nous utilisons la programmation linéaire en nombres entiers pour concevoir des modèles d'optimisation adaptés à la résolution de petites et moyennes instances. Pour traiter les grandes instances, nous proposons des heuristiques en utilisant des techniques de relaxation linéaire.

Outre nos travaux théoriques sur le suivi de la pollution atmosphérique, nous avons conçu et déployé dans la ville de Lyon un réseau de capteurs de pollution économe en énergie. Sur la base des caractéristiques de notre système et des jeux de données de la pollution atmosphérique, nous avons évalué l'efficacité de nos approches de déploiement et d'ordonnancement. Nous présentons et discutons dans cette thèse les résultats d'évaluation de performances ainsi que des lignes directrices pour la conception de systèmes de suivi de la pollution de l'air. Parmi nos principales conclusions, nous soulignons le fait que la taille optimale du réseau de capteurs dépend du degré de variation des concentrations de pollution dans la région de déploiement.

Résumé Long

La pollution de l'air est l'un des principaux problèmes auxquels sont confrontées nos villes développées qui, en essayant de fournir des services publics améliorés, augmentent considérablement l'urbanisation et l'industrialisation et ainsi nuisent à la qualité de l'air. Les effets nocifs de la pollution atmosphérique sur la santé humaine ont été largement établis dans plusieurs études. Selon les estimations de l'Organisation Mondiale de la Santé (OMS), l'exposition à la pollution de l'air est la cause d'environ sept millions de décès par an. En outre, plusieurs études ont montré que l'exposition à de fortes concentrations de polluants atmosphériques pouvait entraîner de nombreuses maladies telles que le cancer, l'asthme et les maladies de la peau. Outre les effets néfastes de la pollution de l'air sur la santé humaine, l'environnement subit également les effets des polluants atmosphériques. On peut citer notamment le changement climatique, les pluies acides et la réduction de la visibilité en extérieur.

La première étape dans la lutte contre la pollution atmosphérique consiste à assurer une surveillance efficace et fine des polluants nocifs. En effet, grâce à un suivi efficace de la pollution atmosphérique, le gouvernement et les collectivités locales peuvent utiliser les données de pollution pour identifier les quartiers où la pollution devrait être réduite. Traditionnellement, la surveillance de la pollution atmosphérique est effectuée par les autorités environnementales régionales et gouvernementales. Ces agences utilisent des stations de mesure conventionnelles équipées de technologies de détection de haute qualité leur permettant d'obtenir des mesures précises des concentrations de pollution de l'air. Cependant, ces stations sont massives, inflexibles et très coûteuses à déployer à l'échelle du quartier. Outre la surveillance traditionnelle basée sur les mesures, les cartes de la pollution de l'air peuvent également être obtenues à l'aide des modèles physiques qui simulent le phénomène de dispersion de la pollution de l'air. Les données d'entrée des modèles physiques sont généralement moyennées dans l'espace et dans le temps, ce qui rend difficile l'estimation précise et fine des concentrations de pollution.

Les limitations des solutions de surveillance traditionnelles ont poussé les industriels à construire une solution plus flexible grâce à l'utilisation de mini capteurs de pollution qui sont moins précis mais plus petits que les stations et surtout moins coûteux. Ces capteurs sont conçus pour être déployés en grand nombre dans une ville donnée et communiquent généralement en sans-fil, formant ainsi ce que nous appelons un réseau de capteurs sans fil (RCSF). Les réseaux de capteurs sans fil présentent de nombreux avantages par rapport aux solutions traditionnelles de surveillance de la pollution atmosphérique. En effet, le rapport coût-efficacité des noeuds capteurs par rapport aux stations traditionnelles permet des déploiements à échelle plus fine et améliore donc la granularité spatiale des méthodes de surveillance actuelles. La granularité spatiale

peut également être améliorée en fixant les capteurs sur des véhicules mobiles grâce à la petite taille des noeuds. Les capteurs peuvent également être fixés sur des drones pour la surveillance de la pollution en 3D, ce qui est une limite importante des solutions traditionnelles.

Déploiement et ordonnancement de RCSF

Dans cette thèse, nous considérons trois applications principales concernant l'utilisation des RCSF pour le suivi de la pollution de l'air : 1) la cartographie en temps réel de la qualité de l'air ; 2) la détection de dépassements de seuils des polluants ; et 3) la correction de modèles physiques qui simulent le phénomène de dispersion de la pollution. Tout en prenant en compte les caractéristiques spécifiques du phénomène de la pollution atmosphérique, nous nous attaquons aux problèmes de déploiement et d'ordonnancement de RCSF.

Déploiement de RCSF : La minimisation du coût de déploiement est un défi majeur dans la conception des RCSF. Le problème consiste à déterminer les positions optimales des noeuds capteurs et des puits de manière à couvrir le phénomène physique et à assurer la connectivité du réseau tout en minimisant les coûts de déploiement. Le coût de déploiement comprend le coût des capteurs et des puits, en plus des coûts opérationnels tels que l'énergie dépensée par les noeuds. Le réseau est dit connecté si chaque capteur peut communiquer ses données à au moins un noeud puits. En raison du caractère NP complet du problème de déploiement de RCSF, la plupart des approches existantes utilise un modèle de détection simple qui suppose qu'un capteur peut couvrir un point de l'environnement si la distance qui les sépare est inférieure à un rayon appelé le rayon de détection. Cependant, cela n'est pas adapté à la surveillance de la pollution car un capteur de pollution détecte seulement les polluants qui arrivent à son contact. La notion de rayon de détection n'est donc pas pertinente dans ce contexte.

Ordonnancement de RCSF : Maximiser la durée de vie du réseau est un problème majeur dans les réseaux de capteurs sans fil qui fonctionnent généralement avec des batteries. La définition la plus utilisée de la durée de vie du réseau est la période de temps pendant laquelle le réseau est opérationnel. Cela signifie que la couverture est assurée (i.e. les exigences de couverture de l'application sont vérifiées) et le réseau est connecté (i.e. chaque capteur est capable d'envoyer ses données à au moins un noeud puits). Plusieurs travaux dans la littérature ont ciblé le problème de la maximisation de la durée de vie du réseau et à différents niveaux de conception : le déploiement, l'ordonnancement des capteurs, l'équilibrage de la charge de communication, l'optimisation du débit de transmission, l'optimisation de la puissance de transmission, le routage, etc. Les travaux existants sur l'ordonnancement des capteurs supposent généralement que les capteurs ont deux modes de fonctionnement : un mode actif où la détection, la communication et le traitement peuvent être effectués ; et un mode veille où le capteur consomme très peu d'énergie. L'ordonnancement consiste

ainsi à ne garder qu'un sous-ensemble de capteurs en mode actif et peut être effectué de manière distribuée ou centralisée.

Contributions de la thèse

Dans cette thèse, nous abordons les trois principales applications de la surveillance de la pollution atmosphérique et nous proposons trois solutions adéquates au problème de déploiement des noeuds capteurs et puits en nous basant sur la programmation linéaire en nombres entiers (PLNE) et la relaxation linéaire. Nous proposons également une approche d'ordonnancement adaptée au cas de la correction de modèles physiques. En plus des travaux théoriques, cette thèse comprend également une partie expérimentale dont l'objectif est de concevoir et de déployer un réseau de capteurs de pollution qui est à la fois low-cost et économe en énergie.

Conception et déploiement d'une plateforme de mesure de la pollution de l'air : Dans le cadre de cette thèse, et suivant les efforts des projets existants utilisant les RCSF pour la surveillance de la pollution atmosphérique, nous avons réalisé une partie expérimentale afin de concevoir une plate-forme de mesure de la qualité de l'air. Dans notre premier prototype, les noeuds capteurs mesurent le dioxyde d'azote (NO_2) en plus de la température et de l'humidité et transmettent les données à une passerelle à l'aide de modules de communication LoRa. Le noeud passerelle communique les données au centre de décision via une connexion 4G. Chaque noeud capteur est alimenté par une batterie lui permettant de fonctionner pendant au moins deux mois grâce à notre conception économe en énergie basée sur des composants à faible consommation. Les capteurs sont également équipés de panneaux solaires afin de prolonger leur durée de vie lorsque leurs batteries sont épuisées. Nous avons déployé notre plateforme dans le centre-ville de Lyon pendant 4 mois à compter de mi-juillet 2018, et les premiers résultats montrent que nos capteurs fournissent de très bonnes mesures tout en étant économes en énergie.

Déploiement de RCSF pour la cartographie de la qualité de l'air : Dans ce travail, nous concevons un modèle de programmation linéaire en nombres entiers pour le problème de déploiement de RCSF afin d'assurer la cartographie de la qualité de l'air. Nous formulons la contrainte de la couverture de la qualité de l'air en nous basant sur des méthodes d'interpolation afin de déterminer les positions optimales des capteurs permettant de mieux estimer les concentrations de pollution aux positions où aucun capteur n'est déployé. Notre formulation de couverture prend en compte la dérive des mesures des noeuds capteurs et l'impact des conditions météorologiques sur la dispersion de la pollution atmosphérique. Nous utilisons le concept de flot pour formuler la contrainte de connectivité, ce qui nous permet de garantir que les capteurs déployés peuvent envoyer leurs données de pollution à au moins un noeud puits. Nous proposons également un algorithme de résolution basé sur la relaxation linéaire et la recherche dichotomique. Enfin, nous évaluons notre proposition en utilisant un jeu de données de la ville de Lyon tout en analysant les résultats de couverture et de connectivité.

Déploiement de RCSF pour la détection de la pollution de l'air : Dans ce travail, nous concevons un modèle d'optimisation pour le déploiement de RCSF pour la détection de dépassement de seuil des concentrations des polluants nocifs. En nous basant sur la modélisation de la dispersion de la pollution atmosphérique et en utilisant l'inventaire des émissions de pollution, nous présentons d'abord la couverture de pollution et la connectivité du réseau de manière indépendante. Ensuite, nous concevons un nouveau modèle d'optimisation dans lequel la couverture et la connectivité sont formulées de manière conjointe en utilisant uniquement le concept de flot. Nous montrons au travers de simulations approfondies que cette formulation conjointe améliore le temps d'exécution du modèle avec un facteur allant jusqu'à 10. Le modèle proposé prend en compte la détection probabiliste des capteurs de pollution et l'impact des conditions météorologiques. Nous avons évalué le modèle sur un jeu de données de la ville de Londres afin d'étudier l'impact des entrées du modèle sur les résultats de déploiement.

Déploiement de RCSF pour la correction de modèles physiques : Dans les deux approches précédentes, nous avons supposé que la surveillance de la qualité de l'air ne repose que sur les mesures de capteurs sans prendre en compte les autres sources de connaissances possibles, telles que la modélisation formelle du phénomène physique de la pollution de l'air. Dans la troisième approche, nous considérons un nouveau cas d'application où le but est d'assurer le suivi de la qualité de l'air par une correction efficace des modèles physiques de la pollution de l'air. À cette fin, nous concevons un modèle d'optimisation basé sur des techniques d'assimilation de données permettant de trouver les positions optimales des capteurs qui corrigent au mieux les estimations des modèles physiques. Ensuite, nous appliquons la relaxation linéaire pour concevoir une heuristique adaptée à notre problème. Nous évaluons cette approche en utilisant le jeu de données de la ville de Lyon utilisée dans la première approche. Nos résultats montrent que l'utilisation de l'assimilation de données permet de surpasser les méthodes de déploiement basées sur l'interpolation.

Ordonnancement de RCSF pour la correction de modèles physiques : Outre nos travaux consacrés au problème du déploiement des capteurs de pollution, nous nous intéressons également à l'ordonnancement de l'activité de mesure des noeuds. Notre objectif est d'assurer le suivi de la qualité de l'air pendant une période maximale en réduisant la consommation d'énergie des sondes de mesure de la pollution atmosphérique. Nous concevons dans cette thèse un modèle d'ordonnancement pour identifier les capteurs qui peuvent être éteints afin d'économiser leur énergie si leurs voisins mesurent des concentrations plus ou moins similaires. Outre l'activité de mesure, nous intégrons également l'optimisation des routes de communication entre les capteurs et le noeud puits. Nous évaluons notre approche d'ordonnancement en utilisant les caractéristiques des capteurs réalisés dans le cadre de notre plateforme de mesure.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this introduction chapter, we first present an overview of the air pollution phenomenon while focusing on its impact on both human health and the environment. We also highlight the main traditional monitoring techniques of air quality and explain why wireless sensor networks (WSN) can improve the current knowledge on the air pollution phenomenon. Moreover, we present a summary of our contributions regarding the WSN deployment and scheduling problems in addition to the design of our air pollution monitoring platform.

1.1 Context of the thesis

1.1.1 Air pollution: a major environmental and health threat

Air pollution is one of the major challenging issues in our developed cities which, by trying to provide enhanced public services, highly increase urbanization and thus road traffic which is considered, together with the effects of industrialization, as the main air pollution sources. The harmful effects of air pollution on human health have been widely established in several studies. According to the estimations of the World Health Organization (WHO), the exposure to both outdoor and indoor air pollution is causing around seven million deaths every year [1]. 600.000 of those casualties belong to children under the age of five, for whom a clean air is necessary to avoid premature death [2]. Moreover, several studies [3] [4] have shown that the exposure to high levels of air pollution concentrations can lead to many diseases like cancer, asthma and skin diseases.

In addition to the harmful effects of air pollution on human health, the environment is also suffering from the impact of air pollutants [5]. Examples include global climate change, acid rains and outdoor visibility. Global climate change, which is causing the average temperature to increase on our planet, is mainly due to air pollution. Acid rain which damages both soil and water making them harmful to both wild animals and fish is mainly due to nitrogen dioxide (NO_2), a gas pollutant that is mostly generated by road traffic [6]. Other pollutants like the fine dust particles that we call particulate matter (PM) contribute to reducing outdoor visibility through the construction of smog.

The first step to tackling air pollution is to ensure an effective fine-grained monitor-

ing of the harmful pollutants. Indeed, thanks to an efficient air pollution monitoring, government and local collectivities can use pollution data to identify the neighborhoods where pollution should be reduced. This was the case for instance of PSEG, an American energy company, which used pollution data to reduce methane concentrations in some American cities [7]. Air pollution data can be also used directly by citizens in order to avoid walking or biking along highly polluted streets for instance [8].

1.1.2 Air pollution monitoring and assessment methods

Measurement stations: Traditionally, air pollution monitoring has been operated by environmental regional and governmental authorities. In France for instance, air quality is managed by regional monitoring observatories like AirParif in the region of Paris and Atmo-AURA in the region of Lyon. Those agencies use conventional measurement stations which are equipped with high quality sensing technologies allowing them to get accurate measurements of air pollution concentrations (see Fig. 1.1). The measurements of the stations are used to construct air pollution maps which define pollutant concentrations in $\mu g/m^3$ or ppb, and this using interpolation methods for instance [9]. The traditional measurement stations being accurate, they are however massive, inflexible and very expensive to be deployed at the scale of the neighborhood. Indeed, according to Fig. 1.1, only 6 stations are deployed in the whole city of Lyon, which makes it very difficult to measure the air quality with a high spatial granularity.

Satellite imaging: To cope with the spatial resolution limitation of reference measurement stations, satellite imagery can be used and allows to get a good spatial granularity [10]. However, satellite-based measurements don't provide a good temporal resolution because of the high financial cost and the complex processing algorithms of satellite systems.

Physical models' simulations: In addition to measurement-based traditional monitoring, air pollution maps can be also obtained using physical models (also called atmospheric dispersion simulators) which simulate the phenomenon of air pollution dispersion. Physical models take as input the locations of pollution sources, the pollutant emission rate of each pollution source and meteorological data in order to estimate the pollutant concentration at a given location [11] (see Fig. 1.1). The input data is usually averaged in space and time making it difficult to estimate with precision and fine granularity the air pollution concentrations.

Low-cost sensors: The limitations of traditional monitoring solutions pushed industrials to build a more flexible solution through the use of tiny pollution sensors which are less accurate but smaller than the stations and most importantly way cheaper (see Fig. 1.1). Those sensors are designed to be deployed in a large number in a given city and usually communicate wirelessly forming, therefore, what we call a wireless sensor network (WSN).

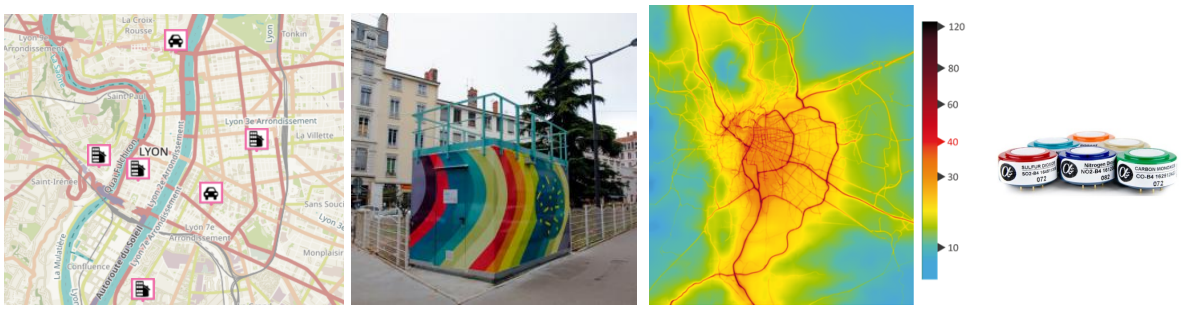


Figure 1.1: Air pollution monitoring and assessment methods. From left to right: (a) the map of the 6 measurement stations located in the Lyon city; (b) a zoom on the station located on the Jean Jaures avenue; (c) nitrogen dioxide annual concentrations ($\mu g/m^3$) simulated using the *Sirane* physical model; and (d) air pollution sensing probes developed by the Alphasense company.

1.1.3 Wireless sensor networks

A wireless sensor network is an ad-hoc network of a set of sensor nodes that are usually deployed in large number in a given environment in order to measure information, and then process it and wirelessly transmit it to a decision center usually via special nodes called sinks (or gateways) [12]. In some cases, the measurements of sensors can be relayed to sinks using specific relay nodes which perform only communication tasks.

The ad-hoc feature of WSN means that the sensor nodes do not need a wireless infrastructure and can self-organize themselves in order to communicate their data to sink nodes. Since sensor nodes are low-cost in order to ensure large deployments, their hardware configuration is limited and their power resources are restricted, which makes the autonomous and self-organizing features of sensor nodes very challenging.

Wireless sensor networks can be applied in several areas which can be grouped into two categories [13]. The first category, named *tracking*, corresponds to the applications where the goal is to track some targets or objects and thus study their evolution in the environment. As for the second category, it corresponds to *monitoring* applications and aims to detect the presence of a target or the evolution of a phenomenon such as pollution, temperature, humidity, etc.

1.1.4 WSN-based air pollution monitoring: benefits and challenges

Wireless sensor networks present many advantages compared to traditional air pollution monitoring solutions [14]. Indeed, the cost-effectiveness of the sensor nodes compared to traditional air quality measurement stations allows for large deployments and hence improves the spatial granularity of the current monitoring methods. Spatial granularity can be also improved by fixing the sensors on top of mobile vehicles thanks to the small size of the nodes compared to the large monitoring stations. Sensors can be also fixed on drones for 3D pollution monitoring which is a big limit of traditional solutions. Moreover, sensor nodes, being small while maintaining a good temporal granularity thanks to their reasonable measuring quality, can be carried directly by citizens to not only perform crowdsensing but also to assess the personal exposure of citizens to air pollution.

In order to take advantage of the promising benefits of wireless sensor networks for air pollution monitoring mainly reliability, cost-effectiveness and energy-efficiency, several scientific challenges need to be addressed:

- **Optimization of the network design:** although the sensor nodes are low-cost and can be deployed in a large scale, the deployment budget is usually limited which means that only the necessary number of nodes should be deployed. As a result, a deployment approach should be used to optimize the number and the locations of sensor nodes while ensuring a reliable monitoring [15]. When the nodes are deployed on vehicles or drones instead of being static, their trajectory should be optimized in order to perform the air quality monitoring with a minimum number of mobile nodes [16] [17]. In addition to the decisions regarding the deployment and mobility optimization, the energy consumption of nodes and mainly sensing probes should be also taken into account in the case of battery-powered nodes. Indeed, the power consumption of pollution sensors is very high compared to other environmental sensors such as temperature or humidity [18]. As a result, an efficient scheduling of both sensing and communication activities of nodes should be taken into account in the design of WSN-based air quality monitoring systems.

- **Data calibration:** as in the case of most of chemical sensing probes, air pollution sensors are very sensitive to environmental conditions, which means that raw data needs to be usually calibrated in order to get correct measurements [19] [20]. Data calibration of air pollution measurements is a very challenging issue not only because of the sensitivity to temperature and humidity, but also to different air pollutants [21]. Indeed, most gas sensors are designed to measure a specific gas pollutant but the presence of other pollutants than the measured one highly impacts the output of the sensing probes.

- **Generation of accurate air pollution maps:** the cost effectiveness of air pollution sensors allows to perform multiple deployments and in different ways: static nodes attached to street light poles, mobile nodes mounted on top of vehicles, etc. [22]. This leads to different data sources which are heterogeneous in both spatial and temporal dimensions and also in terms of sensing quality when using different sensors. As a result, processing all this data to get accurate air quality maps is a big challenge when using wireless sensor networks for air pollution monitoring.

While we focus in this manuscript on two optimization problems of the network design, mainly deployment and scheduling problems, we also present in chapter 3 some of the literature solutions regarding the other scientific challenges that were addressed in the main research projects in the area of WSN-based air quality monitoring systems.

1.2 Thesis challenges

In this thesis, we consider three main applications regarding the use of WSN for air pollution monitoring: 1) the construction of real time air quality maps using sensor

measurements (interpolation-based air quality mapping); 2) the detection of pollution threshold crossings; and 3) the correction of physical models that simulate the pollution dispersion phenomenon. While focusing on those applications and taking into account the specific characteristics of the air pollution phenomenon, we tackle the problems of WSN deployment and scheduling.

1.2.1 WSN deployment

Minimizing the deployment cost is a major challenge in WSN design. The problem consists in determining the optimal positions of sensors and sinks so as to cover the physical phenomenon and ensure the network connectivity while minimizing the deployment cost [15]. The deployment cost includes the cost of sensors and sinks in addition to operational costs such as the energy spent by the nodes. The network is said connected if each sensor can communicate its data to at least one sink node.

Because of the NP-completeness of the WSN deployment problem [23], most existing approaches use a simple detection model which assumes that a sensor can cover a point in the environment if the distance between them is less than a radius called the detection range [24] [25]. The coverage is then modeled as a k-coverage problem in which at least k sensors should monitor each point of interest. This can be true for some applications like presence sensors but is not adapted to pollution monitoring. Indeed, a pollution sensor detects pollutants that are brought in contact by the wind. The notion of detection range is therefore irrelevant in this context. In order to define a realistic formulation of pollution coverage, we propose in this thesis to use pollution propagation models that can take into account the inherently emission rates, weather conditions and their impact on pollution dispersion.

1.2.2 WSN scheduling

Maximizing the lifetime of the network is a major issue in wireless sensor networks which usually operate using batteries [26]. The most used definition of network lifetime is the time period during which the network is operational; this means that coverage is ensured (the application sensing requirements are verified) and the network is connected (every sensor can send its data to at least one sink node). Several works in the literature have targeted the problem of network lifetime maximization and at different design levels: deployment, sensor scheduling, communication load balancing, transmission rate selection, transmission power selection, routing, etc. [27].

Existing works on sensor activity scheduling usually assume sensors to have two operation modes: active mode where sensing, communication and computation can be performed; and sleep mode where the sensor consumes a very small amount of energy [28]. Activity scheduling consists of keeping only a subset of sensors in active mode and can be performed in two ways: 1) in a distributed way where a sensor communicates with its neighbors to decide whether it should turn off or not; or 2) a central way where an optimal sensor schedule is determined by a central node (the sink node for instance).

Without loss of generality, we focus in this thesis on the second case because the air pollution physical models are executed only on the sink nodes due to their high com-

putation requirements. We propose application-aware scheduling models that maximize the network lifetime while ensuring air quality coverage and network connectivity at each time slot.

1.3 Contributions of the thesis

In this thesis, we propose three main adequate solutions to the deployment issue of sensor and sink nodes based on mixed integer linear programming modeling (MILP) and linear relaxation while tackling the three main applications of air pollution monitoring. We also propose a WSN scheduling approach that is adapted to the case of physical models' correction. In addition to the theoretical works, this thesis also includes an experimental part where the objective is to design and deploy an energy-efficient low-cost wireless pollution sensor network.

Contribution 1: Design and deployment of a WSN-based air pollution monitoring platform

As part of this Ph.D. thesis and following the efforts of the existing projects using WSN for air pollution monitoring, we carried out an experimental part in order to design from scratch an energy-efficient air pollution monitoring platform. In our first prototype, sensor nodes measure the nitrogen dioxide pollutant (NO_2) in addition to temperature and humidity and transmit data to a gateway using LoRa communication modules. The gateway node communicates data to the decision center using a 4G connection. Each sensor node is powered using a battery allowing it to run for at least 2 months thanks to our energy-efficient design which is based on low power components and software. The sensors are also equipped with solar panels in order to extend their lifetime when their batteries are drained. We deployed our platform in the downtown of the Lyon city for 4 months starting from mid-July 2018, and the first results show that our sensors provide quite good measurements while being energy-efficient.

The design of the UrPolSens platform allowed us to understand the specific characteristics of air pollution sensors mainly the operation and the energy consumption of the sensing probes. We hence leveraged this knowledge in the design and the evaluation of our application-aware deployment and scheduling approaches.

Contribution 2: WSN deployment for air quality mapping (selected papers: [J2], [C4])

In this work, we focus on the application of air quality mapping where the objective of the monitoring is to construct real-time air pollution maps by interpolating pollution sensor measurements. We tackle the deployment problem and design an optimization model using integer programming modelling. We formulate the constraint of air quality coverage based on interpolation methods in order to determine the optimal positions of sensors allowing us to better estimate pollution concentrations at positions where no sensor is deployed. Our coverage formulation takes into account the sensing drift of sensor nodes and the impact of weather conditions on air pollution dispersion.

We use the flow concept to formulate the connectivity constraint that ensures that the deployed sensors can send pollution data to at least one sink node. We also propose a resolution algorithm based on linear relaxation and binary search. Finally, we evaluate our proposal based on a dataset of the Lyon city while analyzing the coverage and connectivity results.

Contribution 3: WSN deployment for air pollution detection (selected papers: [J1], [C6])

In this work, we design ILP optimization models for the WSN deployment problem while focusing on detecting the threshold crossings of toxic pollutants rather than performing complete mapping of air quality. Based on the pollution dispersion modeling applied on pollution emission inventory and the ILP related works, we first present pollution coverage and network connectivity independently. Then, we design a new efficient optimization model where coverage and connectivity are formulated in a joint way using only the flow concept. The proposed models take into account the probabilistic sensing of pollution sensors and is designed to handle multiple scenarios of weather conditions. We applied the models on a data set of the London city in order to assess the impact of the models' inputs on the deployment results.

Contribution 4: WSN deployment for physical models' correction ([C2], [C9])

In the two previous approaches, we assumed that the monitoring of air quality is based only on the measurements of sensors without taking into account the other possible knowledge sources such as the formal modeling of the physical phenomenon of air pollution. In the third approach, we consider a more challenging application case where the aim is to ensure air pollution monitoring through the effective calibration of the physical models of air pollution. To that end, we design an optimization model based on data assimilation techniques that allow to find the optimal sensor positions which correct in the best way the estimations of the physical models. Then, we apply linear relaxation to design a heuristic algorithm that is adapted to our problem. We evaluate this approach using the data set of the Lyon city used in the first approach. Our results show that using data assimilation allows to outperform interpolation-based deployment methods.

Contribution 5: WSN scheduling for physical models' correction ([C1])

In addition to the works targeting the deployment issue of pollution sensors, we are also interested in the scheduling of the sensing activity of nodes. Our aim is to ensure air quality monitoring for a maximum period by reducing the power consumption of high energy-consuming air pollution sensing probes. We design in this thesis a scheduling model to identify the sensors that can be turned off in order to save their energy in case their neighbors measure similar pollution concentrations. In addition to the sensing activity, we also integrate the optimization of communication routes between nodes, based on a flow model that is adapted to the air quality application. We

evaluate our scheduling approach using the characteristics of our lab-designed pollution sensor nodes.

1.4 Organization of the following chapters

After this general introduction, we present in chapter 2 the state of the art of WSN deployment and scheduling problems while focusing on the most related works to our proposed approaches. Next, we review in chapter 3 the main research projects which focus on the design of WSN-based air pollution monitoring platforms. We also present our lab-designed platform in chapter 3. Then, we present our deployment approach for air quality mapping followed by our deployment approach for the application of air pollution threshold crossings' detection in chapters 4 and 5 respectively. After that, we focus on the application of the correction of air pollution physical models and present our deployment and scheduling approaches in chapter 6. Finally, we conclude the manuscript and discuss some future directions of our works in chapter 7.

1.5 List of publications

In Journals

[J1] Ahmed Boubrima, Walid Bechkit and Hervé Rivano. Optimal WSN Deployment Models for Air Pollution Monitoring. In IEEE transactions on wireless communications, 2017 [29].

[J2] Ahmed Boubrima, Walid Bechkit and Hervé Rivano. On the Deployment of Wireless Sensor Networks for Air Quality Mapping: Optimization Models and Algorithms. In IEEE/ACM Transactions on Networking, 2019 [30].

Book chapters

[B1] Ahmed Boubrima, Walid Bechkit and Hervé Rivano. On the Optimization of WSN Deployment for Sensing Physical Phenomena: Applications to Urban Air Pollution Monitoring. In the Springer's Edited Textbook 'The Philosophy of Mission-Oriented Wireless Sensor Networks', 2019 [31].

In International Conference Proceedings

[C1] Ahmed Boubrima, Azzedine Boukerche, Walid Bechkit and Hervé Rivano. WSN Scheduling for Energy-Efficient Correction of Environmental Modelling. In the 15th IEEE International Conference on Mobile Ad-hoc and Sensor Systems (IEEE MASS 2018), Chengdu, China [32].

[C2] Ahmed Boubrima, Walid Bechkit, Hervé Rivano and Lionel Soulhac. Leveraging the Potential of WSN for an Efficient Correction of Air Pollution Fine-Grained Simulations. In the 27th IEEE International Conference on Computer Communications and Networks (ICCCN 2018), Hangzhou, China [33].

[C3] Ahmed Boubrima, Walid Bechkit and Hervé Rivano. A New WSN Deployment Approach for Air Pollution Monitoring. In the 14th IEEE Consumer Communications & Networking Conference (CCNC 2017), Las Vegas, Nevada, USA [34].

[C4] Ahmed Boubrima, Walid Bechkit and Hervé Rivano. Error-Bounded Air Quality Mapping Using Wireless Sensor Networks. In the 41st IEEE Conference on Local Computer Networks (LCN 2016), Dubai, UAE [35].

[C5] Ahmed Boubrima, Walid Bechkit and Hervé Rivano. Optimal Deployment of Dense WSN for Error Bounded Air Pollution Mapping. In the 12th International Conference on Distributed Computing in Sensor Systems (DCOSS 2016), Washington, DC, USA [36].

[C6] Ahmed Boubrima, Frédéric Matigot, Walid Bechkit, Hervé Rivano and Anne Ruas. Optimal Deployment of Wireless Sensor Networks for Air Pollution Monitoring. In the 24th IEEE International Conference on Computer Communications and Networks (ICCCN 2015), Las Vegas, Nevada, USA [37].

[C7] Ahmed Boubrima, Walid Bechkit, Hervé Rivano and Anne Ruas. Wireless Sensor Networks Deployment for Air Pollution Monitoring. In the 21st International Transport and Air Pollution Conference (TAP 2016), Lyon, France [38].

[C8] Ahmed Boubrima, Walid Bechkit, Hervé Rivano and Lionel Soulhac. Cost-precision Tradeoffs in 3d Air Pollution Mapping using WSN. In the Second International Symposium on Ubiquitous Networking (UNET 2016), Casablanca, Morocco (Best paper award) [39].

[C9] Ahmed Boubrima, Walid Bechkit, Hervé Rivano and Lionel Soulhac. Poster: Toward a Better Monitoring of Air Pollution using Mobile Wireless Sensor Networks. In The 23rd Annual International Conference on Mobile Computing and Networking (Mobicom'17), Utah, USA [40].

Chapter 2

WSN Deployment and Scheduling Literature Review

After having introduced WSN and their major importance for air quality monitoring, we now review the state of the art of the works which focus on the two main WSN challenges that we address in this thesis: the deployment and the scheduling issues. Existing works that focus on these two WSN design problems aim to ensure the coverage of the deployment region in addition to the connectivity of the sensor network while optimizing a given network performance criterion. This criterion is usually the overall financial cost of the sensor network in the case of the deployment problem and the network lifetime for the scheduling problem [41].

Both of WSN deployment and scheduling problems rely on coverage and communication models. Coverage models, also called detection models in most of the state-of-the-art works, are necessary to characterize the coverage of the deployment environment of sensor nodes [42]. As for communication models, they are used to characterize the connectivity between sensor and sink nodes [43].

In this chapter, we first review the preliminary concepts which are necessary to understand the state-of-the-art works. After that, we discuss the existing classifications of the literature deployment and scheduling solutions. We then propose to review the main existing works while classifying them to event-aware and correlation-aware methods. Indeed, sensor networks are mostly used to detect harmful events (also called targets) that might occur in the environment. In addition, sensor networks can be also used to monitor a phenomenon where there might be some correlation in the environment (mainly environmental monitoring applications). Finally, we discuss the main differences between the existing works and highlight the need of novel application-aware WSN deployment and scheduling approaches for the case of air quality monitoring.

2.1 Preliminaries

2.1.1 WSN deployment

The deployment problem (also known as the placement problem) consists in determining the optimal positions where sensor and sink nodes need to be deployed in order to

ensure the coverage of the deployment region (or the environment) and guarantee the connectivity of the network [44]. The environment is considered as covered when each point of interest in the deployment region is monitored by at least one sensor node. In the case where each point of interest is monitored with more nodes, say k sensors, the environment is considered as $k - covered$ [45].

In addition to coverage, a deployment approach should also ensure the network connectivity which means that each sensor node should be capable of sending its data to at least one sink node either directly or through relay nodes [43]. The optimal positions of nodes should be determined with respect to a network performance criterion which needs to be optimized [46]. This criterion usually corresponds to the financial cost of the nodes and their deployment but can also correspond to the network lifetime which has to be maximized by minimizing the energy consumption of nodes.

2.1.2 WSN scheduling

Given a set of already deployed sensors, the activity scheduling problem consists mainly in identifying the nodes which either measure information that is similar to their neighbors, or are not necessary to relay their neighbors' data [47]. Those nodes can therefore be turned off in order to save their energy and hence extend the lifetime of the network [48]. Here, the lifetime of the network usually means the time period during which the network is operational; this means that coverage is ensured (the application coverage requirements are verified) and the network is connected (every sensor can send its data to at least one sink node) [27].

The literature works which focus on sensor activity scheduling usually assume nodes to have two operation modes: active mode where sensing, communication and computation can be performed; and sleep mode where the sensor consumes a very small amount of energy [49]. In order to keep only a subset of sensors in active mode, activity scheduling can be performed in two ways: 1) in a distributed way where a sensor communicates with its neighbors to decide whether it should turn off or not; or 2) in a central way where an optimal sensing schedule is determined by a central node (the sink node for instance) [50].

2.1.3 Deterministic vs. random deployment

The deployment of sensor nodes can be performed in two different ways: in a deterministic way or randomly [51]. Deterministic deployment is used for instance in home automation and health applications where the parameters to be measured must be accurate. For example, health control requires the placement of sensors in precise positions in order to measure temperature, ambient humidity, oxygen level, etc [52]. Sensors are also deployed in a deterministic way in the case of smart cities' applications where the sensor node is usually attached to lampposts and streetlights. As for random deployment, a typical scenario includes the case of sensor nodes which are dropped from an aircraft in a hostile environment. Overall, the decision of doing the deployment randomly or in a deterministic way depends on the network size, the deployment zone accessibility and the application-case [53].

The purpose of a deterministic deployment strategy is to find the exact positions of nodes while optimizing some network performance criteria. However, providing the exact positions of nodes is not possible in the case of random deployment. The aim in this case is instead to find the optimal density of the network to deploy [44]. This is usually achieved by using a probability density to express the placement constraints of the nodes. For instance, sensors that are located near the sink node should be deployed in a dense way because they are the ones which consume the most energy as they play an important role in relaying neighboring sensor data to the sink node [54].

In terms of scheduling, both deterministic and random deployment strategies require an efficient scheduling in order to optimize the energy consumption of nodes and therefore extend the lifetime of the network. Indeed, random deployment may lead to some redundant locations with a lot of sensors. In such a case, turning off some of the deployed nodes would increase the network lifetime in a remarkable way [27]. As for the case of deterministic deployment, the need of efficient sensor scheduling is necessary for monitoring dynamic phenomena where the number of sensors ensuring coverage requirements varies depending on time. A typical example is the air pollution monitoring application.

2.1.4 Communication models

A communication model defines the radio range of nodes and allows to determine the neighborhood of sensor nodes and thus say whether two nodes can communicate with each other or not. The choice of a communication model is very important when designing a deployment or scheduling approach because it impacts the definition of network connectivity: a sensor network may be theoretically connected with respect to some communication models and at the same time not connected at all when considering other communication models [43].

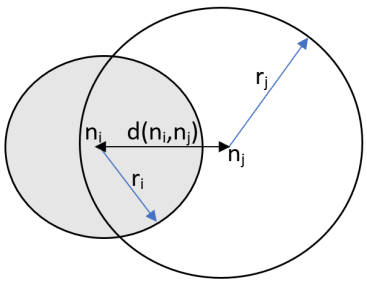


Figure 2.1: Binary disc communication model. Here, n_i and n_j are two sensor nodes and their respective communication ranges are r_i and r_j.

The simplest model that is used in the literature is the binary disc model, which assumes that the communication range of a given node is a disk of radius r called the communication range (see Fig. 2.1). The binary disc model does not take into account the probabilistic nature of communication links. Other sophisticated communication models have been proposed in the literature to cope with this limitation. These models have the ability to take into account the multi-path effects, interference effects, etc [55].

2.1.5 Coverage models

In order to provide a coverage formulation in the context of WSN deployment and scheduling problems, a coverage model (also called detection or sensing model) is required and allows to define the link between the points of interest and sensor nodes [56]. In other terms, a coverage model defines the coverage area of a given sensor node. Most research works on the coverage issue use in fact the simple binary disc model as a detection model and therefore assume that a sensor can cover a point in the environment if the distance between them is less than a radius called the detection range [53]. This can be true for many applications like when using presence sensors for instance but is not adequate for some applications like air pollution monitoring. Indeed, most pollution sensors need to take a sample of the air in order to determine pollution concentrations.

The binary disc model being unrealistic, a probabilistic model can be used instead. The probabilistic detection model is an extension of the binary disc model where the probability of coverage decreases with the distance between the sensor node and the point to be covered [42]. Moreover, the binary and the probabilistic detection models can be combined in order to consider hybrid detection levels as in Fig. 2.2 [57].

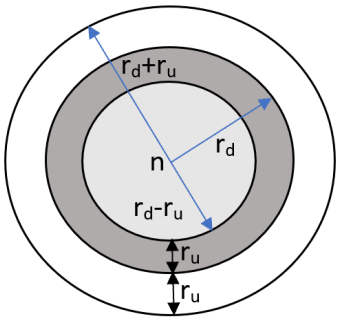


Figure 2.2: Hybrid disc detection model. Here, node n detects all the targets within the smallest range $(r_d - r_u)$. However, the detection is probabilistic if the target is beyond the $(r_d - r_u)$ range and equals to 0 if it is beyond $(r_d + r_u)$.

2.2 Classification of existing works

Both the WSN deployment and scheduling issues have been extensively addressed in the literature where several mathematical models, optimal algorithms and near-optimal heuristics have been proposed [48][58]. The two problems have been defined in multiple ways depending on their context of application. The main issues targeted in the literature of WSN deployment and scheduling are mainly coverage, connectivity, energy consumption and the network lifetime.

2.2.1 Literature classifications

Several classifications have been presented in the literature where the focus was mainly on coverage and connectivity issues regardless whether the aim is to perform WSN

scheduling, WSN deployment or WSN redeployment after an initial random deployment. In [53] for instance, authors consider three different classes of existing works. Works in the first class focus mainly on coverage while providing optimization models based on geometric analysis. Those optimization models formulate coverage while considering the worst-case scenario where a target in the environment is not detected, and the best-case scenario where a given target is detected with a high probability. Formulated this way, the works of the first class are adapted to the case where the objective of the sensor network is to detect intruder targets. In the second class, authors of [53] group the works which assume that a sensor network is already deployed but does not verify coverage requirements. This is the case for instance when sensors are randomly deployed in a hostile environment. In order to improve the coverage provided by the initial deployment, works in the second class leverage sensor mobility in order to relocate the already deployed nodes. In contrary to the first two classes, the third class takes into account the network connectivity constraints in addition to coverage. In most of the works of this class, both the detection model and the communication model rely on the binary disc model. In addition, coverage and connectivity are handled in an integrated way by assuming that the detection range is a multiple of the communication range. This unrealistic assumption makes the works of the third group inadequate to many applications since coverage and connectivity are two different concepts.

A different classification is proposed in [48] where authors divide the existing coverage works into three classes. In the first class, coverage is ensured by monitoring a set of points of interest in a given environment. Those points can correspond to a set of targets to be covered like for instance the missile launchers in a battlefield. Points of interest can also correspond to the result of a discretization of the environment. In this case, the discretization step defines how well coverage solutions are realistic. In order to take into account the continuous nature of the environment in the coverage formulation, other existing works which belong to the second category, called area coverage, use the geometrical characteristics of the deployment region. The third class of works mainly focuses on intrusion detection where the aim is to detect the mobile objects getting in or out of the deployment region. This means that sensor nodes should be located mainly on the perimeter of the environment.

Another classification is proposed in [59] and divides the existing coverage and connectivity works into 4 main classes based on the mathematical approach used to model and solve the problem. Works in the first two classes use, respectively genetic algorithms and particle swarm optimization and hence provide sufficiently good solutions to the coverage and connectivity problems within a reasonable time. In the third class, most existing works borrow some concepts from geometry in order to find the best deployment pattern; i.e. sensors are located following a regular pattern like a triangle or a hexagon [60]. The fourth and last class groups the works which leverage the similarity between mobile robots' control and mobile sensors' deployment. Most of the works of this class borrow therefore concepts from the field of robotics in order to control the mobility of sensor nodes while ensuring coverage and network connectivity. Overall, applying a work from the four classes presented in [59] should be motivated by the nature of the deployment region and the availability of mobility features.

2.2.2 Classification of existing works based on coverage definition

We propose to classify the existing deployment and scheduling approaches based on their coverage definition. Indeed, WSN coverage formulation is a fundamental problem where the complexity varies depending on the application. For instance, when dealing with complex physical phenomena that are difficult to characterize like air pollution monitoring, it is very challenging to define adequate coverage formulations. Within our classification of the literature works, we identify two main groups: event-aware and correlation-aware WSN deployment and scheduling approaches. In the first class, a sensor is assumed to have a detection range, usually circular, within which the sensor can detect any event that may happen. The second class is, however, based on the correlation that sensor measurements may present in order to select the minimum number of sensor nodes and their optimal locations.

2.3 Event-aware deployment and scheduling methods

2.3.1 WSN deployment and the set cover problem

Chakrabarty et al. were among the first authors to tackle the problem of WSN deployment in their works [23][61] where they focus on the case of event (or target) detection. They represent the deployment region as a grid of points which can be two-dimensional or 3D. The authors formulate coverage while assuming that each sensor has a circular detection area, which defines the points that a sensor can cover (see Fig. 2.3). The authors also consider multiple types of sensor nodes with different detection ranges and different costs. However, only coverage is taken into account in their proposal, which means that the optimal positions computed by the proposed solution may result in a disconnected network.

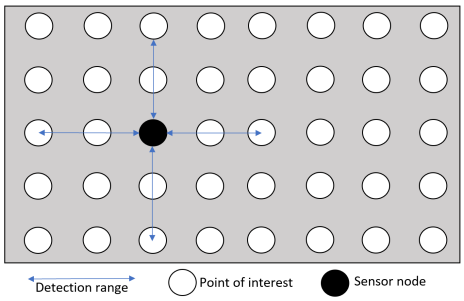


Figure 2.3: Representation of the deployment region in Chakrabarty et al.'s work.

In the first step of the works of Chakrabarty et al., they propose a nonlinear formulation for minimizing the deployment cost of sensors while ensuring a complete coverage of the deployment region. They ensure coverage by forcing each point of interest to be within at least k sensor nodes. In order to be able to use integer programming solvers, they linearize their first formulation using state-of-the-art techniques [62], which allow them to solve medium size instances. They also propose an efficient divide-and-conquer approach in order to deal with large size instances. while providing a good approximation of the optimal solutions.

Since it is based on a nonlinear formulation, Chakrabarty et al.'s model suffers from the intractability; i.e. the formulation is so complex that even small instances cannot be solved in a reasonable time. In addition, only circular detection zones are considered in this work, which is not realistic in some applications. Meguerdichian and Potkonjak [63] dealt with these drawbacks and proposed an ILP formulation of coverage. Compared to Chakrabarty et al.'s model, Meguerdichian and Potkonjak leverage the analogy between coverage and the Set Cover Problem [64], an NP-complete optimization problem. Authors consider sensors which can have a detection area with a shape that is not necessarily circular. Their ILP model does not take into account sensor nodes which have different costs and characteristics and treats only 1-Coverage where a point of interest in the deployment region must be covered by at least one sensor node.

The integer programming formulations proposed in [23] and [63] do not take into account the different coverage requirements of the environment points. Instead, the authors assume uniform coverage where all the points in the sensor field must be covered by the same number of sensors. This may be unrealistic in some applications where some zones of the environment are more critical than the others. Altinel et al. [65][66] focused on this issue and proposed an integer linear programming formulation that considers different coverage requirements among the sensor field points. They also extended their formulation to take into account the probabilistic sensing of sensor nodes while assuming that a node can cover a given point with a certain predefined probability. The authors also propose heuristic algorithms based on their ILP formulation. They show that their approximated solutions have a drift of 11.3 compared to the results of the optimal models. In addition, the authors state that their approach can deal with connectivity under the assumption that the transmission range is as least twice the detection range [67]. In this case, guaranteeing coverage leads automatically to a connected network. However, this assumption is unrealistic for many applications, especially when considering probabilistic communication and detection models.

As in [65] and in order to take into account the case of probabilistic sensing where some errors may occur in sensor measurements, Dhillon et al. proposed in [68] iterative deployment algorithms which are well adapted to large-scale instances. The authors take into account both average and worst-case coverage. In the first case, the objective is to minimize the number of sensors while ensuring that the average number of the non-detected events occurring in the environment does not exceed a given threshold. As for the worst-case coverage, the objective is to reduce the number of non-detected events at each point in the deployment field. The authors compare their deployment solutions to random deployment and show an improvement with a factor going up to 6.

Another work dealing with WSN event-aware deployment under coverage uncertainty assumptions is proposed in [69][70]. This work applies data fusion in the coverage definition in order to take into account the sensors' collaborative detection of targets. The authors based in their work on a probabilistic sensing model in order to define the probability of target detection in addition to false alarm rates (which correspond to the case where an unexciting event is declared by a sensor node). Then, the authors formulated a nonconvex optimization problem minimizing the number of nodes under coverage constraints. They presented resolution algorithms and showed

that the obtained solutions are near-optimal and hence very close to the optimal ones. In addition, and according to their simulation results, their resolution algorithms allow them to get the near-optimal solutions with a speedup factor of up to 7.

2.3.2 Random deployment and redeployment solutions

Unlike the works presented in [23] and [68] where the focus was on deterministic deployment, Zou et al. considered in their works [71][72] the case where sensors are randomly deployed in the sensor field. Even though in such a case sensors are deployed using a well-studied probabilistic density, the obtained positions of nodes do not allow necessarily to get an effective coverage. To cope with that, the authors of [71] propose to redeploy the sensor nodes in order to ensure coverage requirements. They propose a Virtual Force Algorithm (VFA) which uses virtual attractive and repulsive forces between sensor nodes in order to define the way sensors move from their initial deployment location to a new position while improving the coverage of the deployment field. Once the new sensor locations are identified, the nodes move while optimizing their energy consumption due to mobility. Besides the works of [71] and [72], the concept of virtual forces has been widely used in the literature in order to provide an efficient redeployment of sensor nodes. Examples of other existing works include [73][74][75][76][77].

2.3.3 Scheduling strategies and the maximum set covers problem

Having bad coverage is not the only outcome of random deployment. Indeed, in some cases the network may be over-dimensioned meaning that a lot of sensors may be monitoring the same targets even if it is not required by the application. Figure 2.4 shows an application example of the scheduling problem. The environment is represented by three points, and four sensors (S_1, S_2, S_3, S_4) were selected while respecting the budget constraints during the deployment phase of the network. These sensors are scheduled with respect to four time intervals. For instance, S_4 is active only during the last time period.

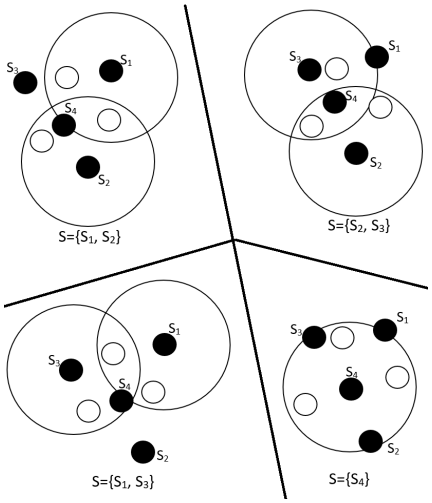


Figure 2.4: An example of the WSN scheduling problem.

Zou et al. tackled the scheduling problem and proposed in [78] to schedule the sensing activity of sensor nodes which would extend the network lifetime by keeping only a subset of sensor nodes activated at a given time period. The scheduling problem defined this way is based in fact on an NP-complete problem called Maximum Set Covers where the idea is to maximize the number of sensor subsets which cover the environment. The authors of [78] use integer linear programming to propose first an optimal centralized solution that is usually executed on sink nodes. In addition, authors also propose a distributed approach and show its efficiency using ns2 simulations.

As in [78], the authors of [79] consider the case where sensors are randomly deployed for target detection. They focus on the problem of finding the maximum number of disjoint sets where each sensor set can ensure coverage. They propose to solve this problem using a hybrid genetic algorithm. In this work, authors don't take into account the connectivity of the network and assume that sensors have the same sensing range and the same initial amount of energy. They justify their choice of not taking into account the network connectivity by the fact that in most target detection applications, a relationship between the sensing range and the communication range of nodes allows them to get a connected network by ensuring only the coverage constraints.

In [80], authors consider the application case of target detection using heterogeneous sensor nodes. They propose to use a heuristic based on ant colony optimization to maximize the number of disjoint connected sensor cover sets and hence get a sensor schedule which maximizes the lifetime of the network. Compared to the work of [79], the authors of [80] take into account both connectivity and nodes' heterogeneity while considering two types of nodes: sensors and relay nodes. In [81], authors focus on the particular case of barrier coverage where targets crossing the border are monitored using sensors that are located on the perimeter of the deployment region. They assume that the already deployed sensor nodes are mobile and propose to maximize the lifetime of the network by minimizing the energy consumption due to both sensing and redeployment mobility. Authors assume in this work that sensor nodes are identical. In addition, they don't take into account connectivity constraints.

In [82], authors also focus on the application of target detection while providing a joint deployment and scheduling optimization approach. They consider the case of heterogeneous WSN nodes and provide a linear mathematical model and heuristic algorithms. Their approach is to optimize the network lifetime given a deployment budget by minimizing the energy that is due to both sensing and data routing. As in [82], authors in [28] focus on both deployment and scheduling for the application case of target detection. They consider a fixed set of sensors to be deployed and scheduled with different sensing capabilities. However, they don't take into account the connectivity of the network as many of the existing works.

2.3.4 Recent works including multi-objective design

Keskin et al. propose in [15] an integer linear program which takes into account the mobility of sink nodes in addition to the formulation of connectivity based on the flow concept while assuming that sensors generate flow units in the network and verify if sinks are able to recover them. They show through simulations that this extends

the lifetime of the network and explain this by the fact that the sensors that are placed around the static sink nodes tend to exhaust their batteries quickly given the concentric nature of communications in WSN. Figure 2.5 shows an application example of the approach of [15]. Here, the two sink nodes, depicted using triangles and dots, move at each time period to allow some sensors to go into sleep mode and thus reduce their energy consumption.

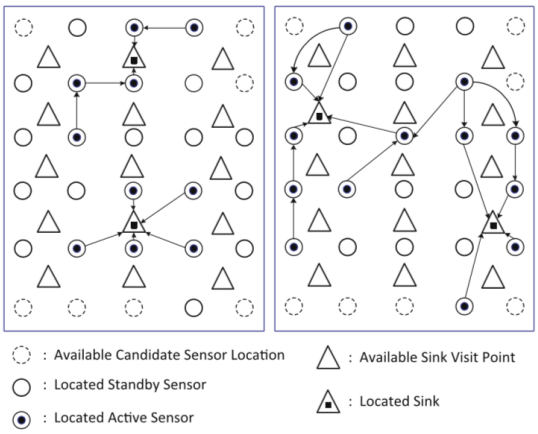


Figure 2.5: An example showing mobile sinks' deployment in Keskin et al.'s work. Here two mobile sinks are used to ensure connectivity. Two time periods are considered, and sinks move back and forth between two locations to gather sensor data.

In terms of connectivity, another formulation has been introduced in [24] where authors base on an assignment approach. They introduce in their ILP model new variables to define the communication paths between sensors and sinks. However, this model involves more variables than the one based on the flow problem and is therefore more complex. In another work [25], authors study the trade-off between coverage, connectivity and energy consumption. They formulate the problem as an ILP model and then propose a multi-objective approach to optimize coverage, the network lifetime and the deployment cost while maintaining the network connectivity.

As in the deployment problem, recent scheduling works take into account more constraints in the design of the optimization models and algorithms. In [83], authors propose a scheduling technique to select not only the sensing nodes to turn on, but they also schedule the communication paths during data collection while balancing the load due to communication. As a result, they improve the network lifetime compared to the approaches which take into account only the scheduling of the sensing activity of sensor nodes. Another interesting extension in the recent WSN scheduling methods include the work of [84] where authors take into account the effective data collection jointly with the scheduling decisions.

2.4 Correlation-aware deployment and scheduling methods

Compared to the state-of-the-art of event-aware deployment and scheduling methods, the class of correlation-aware works such as [85][86][87][88][89] focuses on the appli-

cations which are characterized by inherent correlations that make sensors measure more or less similar values in some particular cases. In [85], Roy et al. tackle the problem of finding the most informative locations of sensors for monitoring environmental applications; i.e. the locations where sensors should be deployed and kept in active mode. They assume the existence of a set of data snapshots characterizing the spatio-temporal variability of the phenomenon to monitor (see Fig. 2.6). Then, they formulate the problem of finding the best locations of sensors to reconstruct the data of the whole phenomenon with a required precision. The authors propose two optimization models in order to handle both stationary and non-stationary-fields. Based on the proposed models, they design an iterative resolution algorithm to perform their simulations, which show that the more the spatial and temporal correlation in the phenomenon distribution, the less is the number of required sensors.

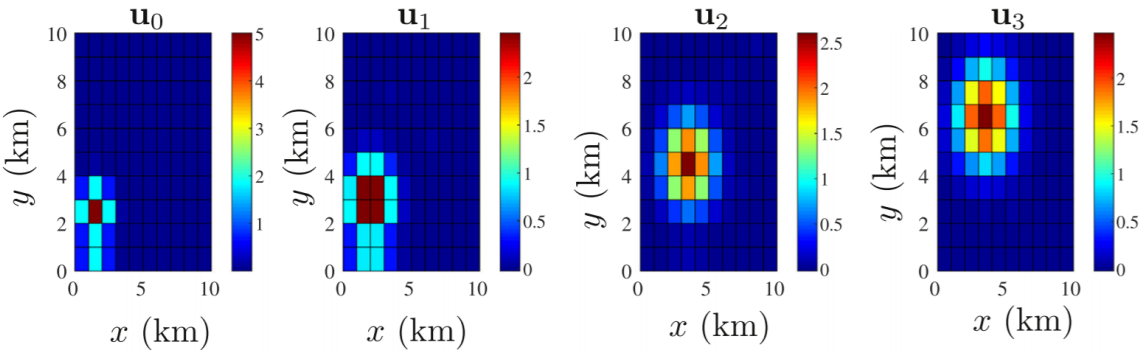


Figure 2.6: An example of the spatio-temporal variability of a generic Gaussian phenomenon. Here, u_i parameters correspond to the time snapshots of the phenomenon.

It is worth-mentioning that the work of [85] is based on the assumption of input data being perfect, which is not the case of several environmental applications where simulated data may present some errors. In [86], Krause et al. tackle the same problem based on the assumption that the variations of the phenomenon are Gaussian (see Fig. 2.7). They also assume a pre-deployment phase allowing to gather data that can be used to characterize the phenomenon. In order to select the best positions of sensors, they use the concept of mutual information in order to define the quality of a given sensor topology. Mutual information is an information theory concept which defines how much information is shared between two random variables. In the context of [86], this allows to evaluate the estimation quality of ground truth values based on sensor measurements while considering both ground truth and measurements as random variables. The work of [86] considers only coverage and is extended in [90] in order to take into account the cost of connectivity while assuming that the communication link quality is Gaussian. Based on extensive simulations, authors show that the criterion of mutual information used in their design allows to provide better estimation of environmental data compared to existing works. However, the works of Krause et al. still assume the phenomenon to be Gaussian, which limits the application of their solutions.

The mathematical characteristics of the correlation-aware WSN deployment case have been studied by Ranieri et al. in [91] while considering only coverage constraints. In order to solve the problem, and based on matrix theory, the authors propose a

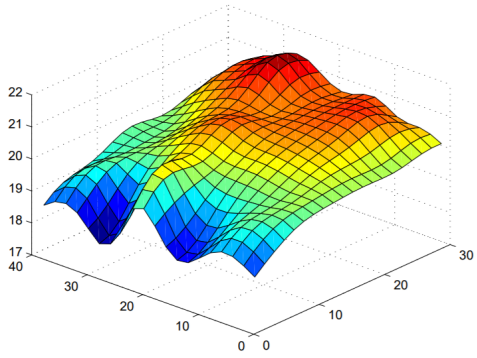


Figure 2.7: An example of Gaussian simulations of temperature variations used as input to Krause et al.'s work.

greedy heuristic algorithm. They show using simulations that their algorithm can outperform existing solutions with a factor of 2 in terms of the estimation error at the locations where no sensor is deployed. However, this work still needs to be adapted and evaluated in order to take into account network connectivity constraints.

In [87], authors tackle the scheduling problem and as in all of the aforementioned correlation-aware works, they focus on field estimation applications where sensors are deployed in order to construct a spatio-temporal field of a given physical parameter. In order to maximize the lifetime of the network, the authors of [87] determine the maximum number of disjoint sensor sets as in the work of [79]. The sensor selection in the scheduling process is based on the data that is gathered by sensors right after their deployment. Authors propose to first learn the characteristics of the physical stochastic process and then use these characteristics as a reference in the definition of the sensing schedule. They show in their simulations that by using their proposal, the energy consumption per node compared to non-scheduled networks can be considerably reduced. For instance, when considering very dense networks, scheduled nodes consume down to 91% less energy. It is noteworthy that this work makes a strong assumption, that is the stochastic process describing the physical phenomenon is stationary. In addition, even though this work focuses only on coverage, minimizing the number of active sensors would lead to less packets being sent to the sink node, which minimizes the energy consumption due to connectivity. However, authors don't take into account the connectivity constraints between active nodes, and the sensing capabilities of nodes are assumed to be homogeneous.

Different from many of the existing scheduling works, the authors of [92] propose to consider the case where nodes monitor not only one physical parameter but several ones and at the same time. This means that when a node is scheduled to turn-off, all the involved sensing probes should be taken into account. Authors propose both centralized and distributed algorithms in order to determine disjoint sensor sets where each set ensures the coverage of all of the monitored physical parameters. Authors perform extensive simulations and show that their solutions can improve the network lifetime with a factor of 6 compared to random deployment. However, authors don't take into account the connectivity of the network, which should have reduced the network lifetime in their simulation results.

2.5 Summary and comparison

We summarize in table 2.1 the main characteristics of a selection of the state-of-the-art WSN deployment and scheduling existing works. We highlight the case where the literature works solve the problem of WSN deployment or WSN scheduling or both. We also indicate when the works take into account the main component of the network, which is sensor-to-sink connectivity, in addition to their ability to adapt to the heterogeneous case of sensor nodes.

- First of all, we notice that most of the existing works fall into the category of event-aware methods and only few of them are adapted to the correlation-aware case. We recall that the sensing (or detection) range of sensor nodes is the common characteristic between all the works of the event-aware class.

- As shown in table 2.1, only few works tackle the problems of WSN deployment and scheduling at the same time in a joint way. Indeed, both problems are NP-complete and even when solving them separately, the execution time is exponential which makes it very difficult to propose polynomial-time near-optimal algorithms to the joint deployment-scheduling problem.

- Many WSN deployment and scheduling existing works don't take into account the connectivity of the network in the design of the optimization models and algorithms. In fact, in those works the connectivity constraint is relaxed based on the assumption that coverage involves connectivity for some applications where the detection range is small compared to the communication range of the nodes. Nevertheless, this is not the case of many WSN applications.

2.6 Discussion

The most recent WSN deployment and scheduling works provide efficient optimization models and algorithms while taking into account realistic network characteristics like the probabilistic nature of communication links and the energy-efficient scheduling of sensor-to-sink communication paths. Moreover, the most sophisticated works provide efficient algorithms for both the deployment and scheduling problems in a joint way, and in some cases even other WSN design issues are taken into account like the mobility of sink nodes for a better data collection. However, the coverage modelling in most of the literature works is generic and not well adapted to some specific applications like air pollution monitoring. Indeed:

- The WSN deployment and scheduling methods which fall into the category of event-aware works are all based on detection models. Whether it is a deterministic detection model or a probabilistic one, this is not adequate for air pollution monitoring because most pollution sensors need to get into contact with the pollutants in order to measure their concentrations.

Author(s)	Work class	Deployment	Scheduling	Connectivity	Nodes' Heterogeneity
Chakrabarty et al. [23]	Event-aware	Yes	-	-	Yes
Altinel et al. [65]	Event-aware	Yes	-	-	Yes
Rebai et al. [24]	Event-aware	Yes	-	Yes	-
Sengupta et al. [25]	Event-aware	Yes	-	Yes	-
Hu et al. [79]	Event-aware	-	Yes	-	-
Liaskovitis et al. [87]	Correlation-aware	-	Yes	-	-
Lin et al. [80]	Event-aware	-	Yes	Yes	Yes
Du et al. [81]	Event-aware	-	Yes	-	-
Keskin et al. [15]	Event-aware	Yes	Yes	Yes	Yes
Mini et al. [28]	Event-aware	Yes	Yes	-	Yes
Chen et al. [83]	Event-aware	-	Yes	Yes	-
Deng et al. [92]	Correlation-aware	-	Yes	-	-
Lu et al. [84]	Event-aware	-	Yes	Yes	-

Table 2.1: A comparison between a selection of the state-of-the-art WSN deployment and scheduling works.

- As for the correlation-aware works, a strong assumption like sensor data being Gaussian is usually considered in the literature making it difficult to derive realistic deployment and scheduling results in our application case. Indeed, the dispersion of pollutants in the air is more complicated than the Gaussian distribution and is usually handled using physical models which take into account the impact of weather conditions and pollution emissions.

Novel application-aware deployment and scheduling methods have been recently proposed in the literature in order to take into account the characteristics of the application case in the design of the optimization models and algorithms; examples include the works of [93][94] on wind monitoring, the works of [95][96] for hot server detection in data centers and the works of [97][98][99] for water pollution monitoring. Following the same direction, we propose in the next chapters to consider the context of air pollution monitoring and present appropriate formulations of coverage in addition to network connectivity.

2.7 Conclusion

An effective deployment and scheduling of wireless sensor networks is necessary to guarantee the application requirements in terms of coverage and connectivity. We reviewed in this chapter the main existing works while classifying them into event-aware and correlation-aware solutions. The largest part of the deployment and scheduling literature works focus on the application of event detection and is based on detection models that define the way targets are detected by sensors. The use case of our thesis, which is the air quality monitoring, is closer to the second part of the literature works which are based on the correlation that might exist between sensor measurements. Even though the works of this second category can be used for some environmental applications like temperature and humidity monitoring, still those works are not well adapted to the application case of air pollution monitoring. Our aim in the following chapters is therefore to present application-aware deployment and scheduling methods and this based on the characteristics of the air pollution phenomenon.

Chapter 3

Design of WSN-based Air Quality Monitoring Systems

The optimization of the deployment and scheduling of pollution senor nodes cannot be effective if the nodes of the sensing platform are not themselves energy-efficient and cost-effective. In order to target this issue, we designed as part of this thesis a low-cost WSN-based air quality monitoring platform while focusing on the energy consumption of nodes compared to existing platforms.

In this chapter, we first review some of the main WSN platforms which have been designed by both researchers and industrials for the application of air pollution monitoring. We present the main features of these monitoring systems, their architecture and the results of their deployments. We also discuss the scientific challenges that have been raised by researchers based on the deployment results.

The second part of this chapter is dedicated to the air quality monitoring system that we designed from scratch and deployed in the Lyon city as part of the UrPolSens project. We highlight the main characteristics of our system compared to existing platforms and mainly in terms of cost-effectiveness, energy efficiency and modularity which make our platform viable in multiple deployment scenarios. We also present the architecture of our system and the deployment that we have carried out in the Lyon city while discussing the first obtained results. We show that our system provides quite good measurements while being energy-efficient.

3.1 Literature review of existing platforms

In this section, we present a selection of the main WSN-based air quality monitoring systems designed as part of some of the most successful research projects.

3.1.1 OpenSense platform in Zurich and Lausanne

OpenSense [100] is one of the most leading research projects which show the emerging potential of wireless sensor networks for air quality monitoring. Taking place in both Zurich and Lausanne cities in Switzerland, the project aims to design, deploy and study the performance of measurement nodes equipped with low-cost air pollution

sensing probes. These sensing nodes are designed to be mounted on public transportation vehicles mainly trams and buses. They can be also fixed on a city infrastructure such as lampposts and traffic signal poles provided that a power supply is available. In addition to temperature and humidity sensing, the OpenSense platform measures several pollutants mainly particulate matter (PM), ozone (O_3), carbon monoxide (CO) and nitrogen dioxide (NO_2). In order to geolocate measurements, the sensor nodes are equipped with a GPS module in addition to an accelerometer which is used to reduce the uncertainty of GPS measurements. Indeed, when mounted on trams for instance, the accelerometer allows to take into account the position of the tram stops in the geolocation process. The core part of the sensor node, which runs under the Linux operating system, is connected to a data center using either WiFi or cellular network. Data is also stored locally at the measurement nodes in case of connectivity issues. Based on the data gathered by the data center, the OpenSense platform provides a visualization interface which shows the sensor measurements on top of the map of the city. In addition, pollution data is also used to help reduce the impact of air pollution on citizens. Indeed, as part of the monitoring platform, a mobile application is proposed to assist pedestrians and cyclists in Zurich in finding the healthiest route from a point A to a point B while avoiding exposure to high air pollution concentrations [8].

The main deployment of the OpenSense platform took place in Zurich where 10 measurement nodes have been mounted on mobile trams in addition to a static node fixed next to a reference station. The aim of using the static sensor is to help in the calibration of the mobile nodes by comparing sensor measurements to the data of the reference measurement station. Offering one of the largest air pollution datasets, the OpenSense sensor network had been operational from 2012 to 2017 in a region of $100km^2$. Because sensor nodes do not operate on battery, they are activated only during daytime and are turned off when the trams are in their depots. In addition to the Zurich deployment, the OpenSense platform is also present in the city of Lausanne where sensor nodes are mounted on top of 10 buses.

The pollution dataset collected using the OpenSense platform allowed studying one of the main challenges of WSN-based air quality monitoring systems which is the effective calibration of low-cost air pollution sensing probes [22]. Indeed, the measurements of low-cost gas sensors are very dependent to environmental conditions in addition to being cross-sensitive. The cross-sensitivity of gas sensors means that for instance an NO_2 sensing probe may capture the variations of the O_3 pollutant which makes the NO_2 measurements erroneous. Researchers within the OpenSense project propose to collocate sensing probes of different pollutants and then to use multiple regression in order eliminate the cross-sensitivity errors [19]. The proposed technique allowed the authors to reduce the calibration error of the measurements by up to 45% compared to state-of-the-art calibration techniques which do not take into account the effects of cross sensitivity.

Another challenge raised by the researchers of the OpenSense project is the need to regularly calibrate mobile nodes because of the sensing drift that occurs over time. They propose to perform multi-hop calibration of the mobile nodes against the measurements of static reference stations [20]. The idea is to first calibrate the nodes which pass near the static reference stations and then consider those calibrated mobile nodes as virtual reference stations. The multi-hop calibration is performed on rendezvous

points where the virtual reference stations meet the other mobile nodes. Compared to state-of-the-art techniques, the multi-hop calibration allows the authors to reduce the calibration error by up to 60%.

OpenSense researchers also worked on the optimization of the sensor network while focusing on the problem of selecting the vehicles of the public transportation system on which the sensors should be mounted. Here, the objective is to cover the maximum part of the city of Zurich for instance with a minimum number of trams [16]. The proposed optimization technique also ensures an effective calibration of the mobile nodes by allowing the mobile vehicles to be able to meet each other from time to time in order to perform the multi-hop calibration process.

3.1.2 Mobile sensor networks using Google Street View cars

Leveraging the availability of public transportation vehicles allowed the OpenSense platform to cover a big part of Zurich and Lausanne cities while improving both spatial and temporal resolutions of air quality maps compared to static reference stations. However, the coverage of the monitoring region can be further improved by transporting sensor nodes using dedicated cars whose routes are not restricted as in the case of buses and trams. In this case, the speed of the cars can be also adapted in order to slow down and hence reduce the effects of the airflow which may impact the quality of sensor measurements. Motivated by these reasons among others, Google launched a series of projects in partnership with the American organization Environmental Defense Fund (EDF) and the Aclima company in order to design and deploy an air pollution monitoring system mounted on Google Street View cars [101]. While using sensing probes that are more accurate than the ones used in OpenSense, the Google platform allows monitoring gas pollutants such as nitric oxide (NO), nitrogen dioxide (NO_2) and methane (CH_4) in addition to particulate matter pollutants such as black carbon (BC). In addition, sensor nodes are equipped with GPS modules for measurements' geolocation and use cellular network to upload pollution measurements to the data center which embeds a real time visualization interface.

Several deployments of the Google monitoring platform took place in different American cities. For instance, 2 Google cars have been used in Oakland in order to ensure air quality mapping of a region of 30 km^2. The measurements took place during weekdays at daytime hours between June 2015 and May 2016. By sensing pollution concentrations literally every second and driving the cars at normal speeds, the Google platform allows a fine spatial resolution in the order of 30 meters. Different from the Oakland deployment, the Google system has been also deployed for methane mapping in order to locate the leaks from natural gas lines in Boston, Indianapolis and New York city [7]. Later, the obtained dataset helped the PSEG energy company replace the defective gas lines.

Among the research challenges facing the Google monitoring system, researchers highlight the fact that in order to identify the difference in pollution patterns from a neighborhood to another, there is a need for an effective driving plan. Such a plan should allow covering every neighborhood of the city regularly and at different times of the day, the week and the year [101]. Researchers also highlight the case of unreliable data in some situations like when a Google car is driving behind a truck in a given

street, which impacts sensor measurements. To fix this issue, they propose to drive several times along the street in question and keep only the median concentration. This considerably reduces the errors of the measurements according to their results.

3.1.3 CityScanner platform in Cambridge, Massachusetts

Even though the Google air quality monitoring system improves the spatial coverage of the monitoring region compared to the OpenSense platform which uses mainly trams and buses, maintaining dedicated monitoring cars can be very expensive and hence limits the number of nodes. As a result, this also limits the temporal resolution of air quality maps. To cope with that, the CityScanner platform [102] designed by MIT researchers and engineers provides a portable air quality monitoring solution which is modular in the sense that it offers many configurations in terms of sensing probes and do not require any specific type of power supply or batteries. Therefore, sensors can be mounted on any vehicle without impacting its operation and independently from whether the vehicle has nonscheduled or scheduled routes. Each sensor node that is mounted on a mobile vehicle includes several sensing modules which communicate raw measurements to a main core component using WiFi. The core component ensures the functions of power management and data storage in addition to GPS geolocation. The core component also communicates time stamped data to a central server using cellular network. In addition to measurement nodes, the CityScanner platform provides an application layer at the central server which aggregates and displays data in real time.

The main deployment of the CityScanner monitoring system took place in the city of Cambridge, Massachusetts where sensor nodes have been mounted on a set of garbage trucks for 8 months. The garbage trucks' deployment allowed covering the entire city during every single week of the duration of the deployment case study. The main purpose of CityScanner deployment was to assess the reliability of the proposed system in terms of energy efficiency, data communication and data accuracy while focusing on testing some sensing modules such as particulate matter, temperature and humidity sensors [18]. The Cambridge deployment allowed researchers to study some of the main challenges that occur when using mobile nodes for air pollution monitoring. Indeed, they highlight the fact that raw data needs to be filtered before being displayed on the visualization interface because of the effects of mobility on the quality of measurements. Researchers also point out the need for efficient data compression methods at the level of sensor nodes because the sensing frequency of mobile nodes is very high due to spatial resolution requirements. This means that a huge amount of data is communicated through cellular network, which also increases the cost of the monitoring solution because of the high cost of cellular communications. This also leads to more data loss and more energy consumption.

3.1.4 AQNet and Astro platforms using drones

By making air pollution sensors mobile, projects like OpenSense, Google's and CityScanner succeeded in improving the spatio-temporal resolution of traditional air quality monitoring systems. Yet, the gathered data allows characterizing pollution variations

only in 2D since sensors are usually mounted at the same height. Researchers from Peking University took another direction in the design of their air quality monitoring system named AQNet [103] where they make use of drones in order to perform 3D mapping of pollution concentrations. In addition to sensors mounted on drones, the AQNet platform uses a set of energy-efficient baseline ground sensors that already cover the deployment region in 2D. Hence, aerial nodes are mainly used to characterize the vertical distribution of air pollution concentrations.

While focusing on particulate matter sensing, the hybrid aerial-ground monitoring system proposed in AQNet has been deployed in the campus of Peking University in Beijing with 200 ground sensors and one aerial node. The obtained dataset helped AQNet researchers study the performance of using neural networks to generate fined-grained air quality maps based on pollution sensor measurements and official data of weather conditions. Moreover, researchers propose in [17] to use the collected measurements to control the operation of drones. Their idea is to first characterize the distribution of air pollutants based on the already collected data. Based on this characterization, they design algorithms to find the positions that need to be visited in order to estimate with high accuracy pollution concentrations in the whole 3D deployment region.

As in the case of the AQNet platform, Astro [104] is a recent and promising project which uses drones to monitor hazardous gas emissions mainly volatile organic compounds (VOCs) that might be released by industrial plants and chemical refineries in the city of Houston, Texas. Compared to AQNet aerial nodes, the drones in Astro do not require ground control stations to guide their operation. In addition, the coordination between the drones to assign sensing tasks is performed in a distributed way thanks to online, light-weight and efficient data analysis algorithms implemented and incorporated directly into the drones' boards. The sensing drones are also energy efficient thanks to the use of software-defined radios which allow to dynamically adapt the communication range of drones and hence reduce the energy consumption that is due to network connectivity.

3.1.5 W-Air platform in Zurich

Differently from the previously presented projects, the W-Air platform [105] which is designed by the same researchers of OpenSense investigates the use of the crowdsensing paradigm in order to involve citizens in the monitoring process. By allowing citizens to carry air quality sensors as wearable devices while moving in the city, the W-Air platform not only improves the spatial and temporal resolutions of air quality maps but also allows citizens to determine their personal exposure to air pollutants. Indeed, crowdsensing nodes measure the quality of the same air that is inhaled by users. The first prototype of the W-Air platform focuses on monitoring carbon dioxide (CO_2) and ozone (O_3) pollutants while embedding temperature and humidity sensors. W-Air measurement nodes use Bluetooth in order to send raw data to smart phones which calibrate measurements and provide a personal visualization interface. Crowdsensed measurements are then sent to a cloud server which performs a second step of data calibration by comparing and analyzing measurements of different users.

In order to evaluate the effectiveness of the monitoring platform, W-Air researchers

performed a set of experimentations while considering 3 different settings of the measurement nodes: mounted on a wristband like a smart watch, attached to a backpack or fixed on a belt. Experimentations took place both indoors and outdoors during 21 days from April to October 2017 while considering different times of the day and different weather conditions. The obtained dataset allowed W-Air researchers to highlight the fact that human beings behavior interfere with the operation of air pollution sensing probes. Indeed, human beings can release multiple natural gases through their skin, textile and exhaling, which may impact the measurements of the sensor device. It is noteworthy that the interference is not restricted to the case of wristbands but also occurs for the other settings like attaching the sensors to backpacks. Therefore, human-related gases need to be filtered in order to get accurate pollution measurements. To that end, W-Air researchers show that the sensing error of measurement nodes can be considerably reduced using a calibration method which is based on neural networks and leverages the fact that human interference is highly correlated with the concentrations of volatile organic compounds ($VOCs$).

3.1.6 Other systems and research studies

The BikeNet platform [106] is another successful air quality monitoring system and uses CO_2 pollution sensors which can be deployed on bikes together with a GPS and an accelerometer. The platform allows users to determine the roads which are the most adequate to biking in terms of both traffic quality and air quality. This helps users improve their cycling experience while monitoring air quality in the city. Airsense [107] is another air quality crowdsensing platform which, in addition to integrating particulate matter sensors, uses contextual sensing information (such as GPS location, whether the user is sitting, standing or walking, etc.) in order to better characterize the exposure of users to air pollution. Indeed, by combining both pollution measurements and contextual information, the platform allows users to determine in which life scenario they are the most exposed to air pollution. For instance, the monitoring system can determine if a user is exposed or not to air pollution when he goes walking in the morning to his work place. While Airsense focuses on particulate matter measurements, other crowdsensing projects like Comon Sense [108] and CitiSense [109] provide monitoring solutions for mainly gas pollutants.

Different from the projects aiming to design air quality monitoring platforms, the AirMonTech European project [110] focuses on the evaluation of the emerging air pollution sensing techniques and provides an open access database [1] in order to guide the decisions regarding the future monitoring technologies. CiteAir [111] is another European project and provides a data visualization platform [2] of air quality in major European cities. The visualization tool compiles data from different monitoring systems across Europe using a common air quality index that is calculated based on the concentrations of the different pollutants. In the same direction of AirMonTech and CiteAir, the MESSAGE project [112] investigates the accuracy of current low-cost pollution sensing technologies in mapping air quality. To that end, experimentations on

[1] http://db-airmontech.jrc.ec.europa.eu
[2] https://www.airqualitynow.eu/

both static nodes attached to lampposts and mobile sensors carried by pedestrians, cyclists and drivers have been performed in the city of Cambridge, UK.

3.1.7 Summary and comparison

We summarize in table 3.1 the characteristics of the main existing WSN-based air quality monitoring platforms. We highlight their main deployment infrastructure, the main monitored pollutants and the communication technologies that they use. In addition to monitoring several gas pollutants, we notice that particulate matter (PM) are the focus of most of existing projects. The main motivation behind that is the fact that those pollutants have been proven to be one the causes of cancer [1].

In terms of communication, most existing platforms tend to use either WiFi or cellular communication. However, it appears that communication solutions that are adapted to the Internet of Things platforms like LoRa and Sigfox are yet to be used in the area of air quality monitoring.

Monitoring platform	Deployment Infrastructure	Main supervised pollutants	Communication technology
OpenSense	Trams & buses	PM, O_3, CO, NO_2	WiFi and cellular
Google's	Google Street View cars	PM, NO, NO_2, CH_4	Cellular
CityScanner	Garbage trucks	PM	WiFi and cellular
AQNet	Drones	PM	Not mentioned
Astro	Drones	VOCs	–
W-Air	Citizens	CO_2, O_3	Bluetooth, WiFi and Cellular

Table 3.1: Summary of a selection of the state-of-the-art WSN-based air quality monitoring platforms.

In addition to leveraging the mobility of sensor nodes, most platforms tend to combine different sensing paradigms in order to improve both spatial and temporal granularity [113]. Static sensing, mobile sensing using vehicles or drones and crowdsensing allow different levels of spatial and temporal granularity. Indeed, as illustrated in Fig. 3.1, sensor mobility when using vehicles or in the case of crowdsensing improves the spatial coverage of air quality maps compared to static monitoring solutions. However, this comes at the cost of the temporal resolution which is reduced because in the case of mobile nodes, a given location is monitored only when a sensor passes by.

3.2 The UrPolSens monitoring system

UrPolSens is a multidisciplinary project which focuses on the use of low-cost wireless sensor networks for fine grained air quality monitoring. The UrPolSens project gathers different research teams from different backgrounds including mainly computer networks, fluid mechanics and sociology. The multidisciplinary consortium of our project allowed us to target three main objectives: 1) the design and the deployment

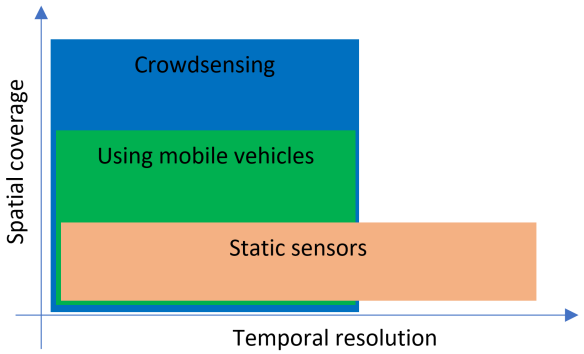


Figure 3.1: Comparison between air quality monitoring paradigms in terms of overall spatial coverage and temporal resolution of pollution maps.

of a WSN-based air pollution monitoring platform which we present in this section; 2) the design of efficient application-aware WSN deployment and scheduling models which we present in the following chapters; and 3) the conduction of a sociological study regarding the air quality perception by citizens.

The sociological study has been carried out together with our sociology partners in order to better understand how most people evaluate their perception of air quality when the air is polluted. We conducted both a main qualitative study and an additional quantitative study in order to determine whether the access to accurate information on air quality mapping would allow people to change their perception of air quality. Indeed, when trying to perceive the quality of the air with our senses, we might sometimes be wrong due to the odorless and invisible characteristics of some pollutants.

In the following of this section, we present our air quality monitoring system while focusing on the objectives of our project compared to existing platforms. We also present the system architecture and the first results of our deployment in the Lyon city.

3.2.1 Objectives and decision guidelines

Our main objectives and decision guidelines regarding the UrPolSens platform are as follows:

- We mainly focus on gas sensors which is a request of our fluid mechanics and sociology partners. However, our platform design remains compatible with PM sensing probes. In our first prototype, we focus on the road traffic pollutant Nitrogen Dioxide (NO_2).

- We also aim to design a cost-effective platform to allow large deployments of sensor nodes by using low-cost sensing probes. After reviewing the NO_2 sensing probes that are available in the market [22], we chose to go with the Alphasense company. Alphasense gas sensing probes have shown good reliability according to many studies [114] and research platforms like W-Air. However, raw data provided by those sensing probes needs to be calibrated against reliable sensors

before the deployment phase. For this purpose, we use a set of medium-level quality sensors *Cairsens*, which are provided by the Cairpol company [115].

- We also aim to design energy-efficient pollution sensors in order to allow long-term deployments and reduce maintenance costs. Different from many existing platforms like Opensense, we build our sensors' board on low-power microcontrollers rather than ready high-energy consuming electronic boards like Arduino and Raspberry Pi. Regarding the communication between nodes, and different from most of existing projects which use either WiFi or Cellular high energy-consuming communication, we use low-power LoRa communication modules. This allows us to get large communication ranges with low-power consumption. In addition, we design efficient software routines in order to reduce the power consumption of the different components of our nodes while using an efficient duty cycling.

- As in the CityScanner platform, we aim to design measurement nodes that are modular and can accommodate different and multiple air pollution sensing probes and environmental sensors. To that end, we build our sensors' circuit board based on Atmega microcontrollers that allow us to connect both digital and analog sensing probes. Therefore, sensing probes delivering measurements using both analog voltages or USB connections are compatible with our platform. Compared to the CityScanner platform, we provide a solution that is modular and energy-efficient at the same time.

3.2.2 Architecture of the platform

We present in Fig. 3.2 the architecture of our platform which consists of three main layers that ensure data collection, storage and presentation. The data collection layer is represented by the measurement nodes which are deployed to monitor air pollutants and weather conditions like temperature and humidity. Sensors can be fixed at static locations but are also designed to fit the mobile monitoring scenarios.

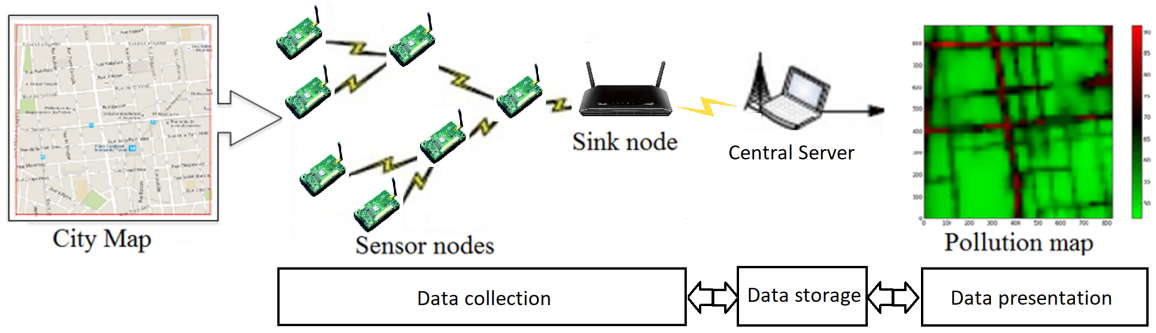


Figure 3.2: Main architecture of the UrPolSens platform.

To ensure air quality monitoring with a fine temporal granularity, sensor nodes perform air pollution measurements every second and then store raw data on their local memory space. At the end of each set of measurements (each 10 minutes for instance),

sensors send their average data to a sink node (or a gateway) using LoRa communication technology. Our main architecture supports multi-hop communication which allows to reach the gateway from a node that is located far away.

Using cellular network, the data received by the sink node is then communicated in real time to a central data server which hosts the database containing all the sensor measurements. In addition to the data storage function, the main server filters and processes raw data in order to show on a web interface real time calibrated measurements of air pollutants and weather conditions. The main interface of our platform is depicted in Fig. 3.3.

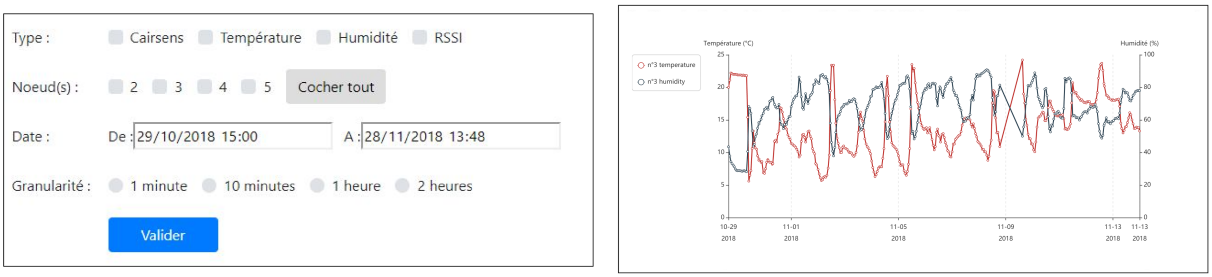


Figure 3.3: Main web interface of the UrPolSens platform. From left to right: (a) The interface data form. Users can display graphics data by measurement node, sensing probe, time period and 1min-to-2h temporal granularity; (b) Filtered and averaged data returned by the visualization interface. Displayed measurements correspond to temperature and humidity data of sensor number 3.

3.2.3 Design of sensor nodes

We depict in Fig. 3.4 the main components of our sensor nodes. Each sensor node consists of three parts: a core module, a set of sensing probes and a power module. The core part of the node performs data processing, storage and communication using, respectively, an Atmega microcontroller (MCU), a short-term EERPOM memory combined with a long-term SD module, and a LoRa radio module. The core part also contains a real time clock (RTC) for data timestamping and an analog to digital converter (ADC) to allow compatibility with analog sensing probes. Digital sensors are supported through the use of the UART protocol and serial communication. Both the core module and the sensing probes are powered using a specific module which combines solar energy and battery usage while switching efficiently between the two power sources. In addition to the components depicted in Fig. 3.4, a GPS and an accelerometer can be added when using the nodes for mobile monitoring.

MCU unit

We use as a MicroController Unit (MCU) the ATMEGA328P-PU which is an 8-bit low-power microcontroller operating on 5V. The MCU hardware design includes several timers that we use to optimize the duty cycling of our measurement nodes. When set in idle mode, the ATMEGA328P-PU draws less than 1mA current. In addition

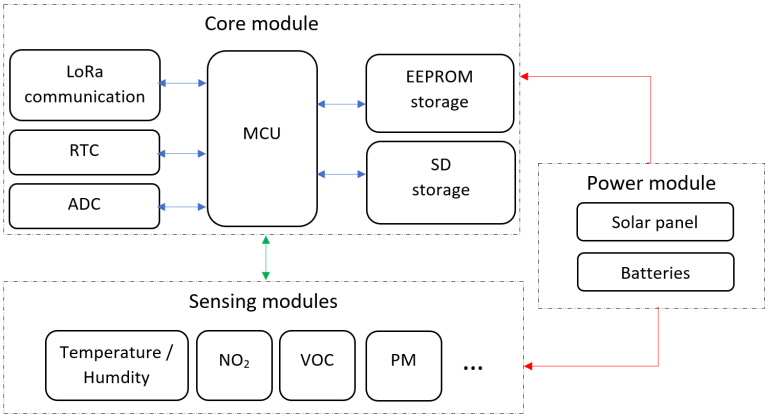


Figure 3.4: Architecture of the UrPolSens nodes.

to its low-power consumption, and combined with a 16-bits high precision ADC, the ATMEGA328P-PU allows us to provide a solution that is compatible with most radio, storage and sensing modules thanks to the availability of multiple communication protocols mainly UART, I2C and SPI.

Raw data storage

Fine-grained raw data is stored locally in an SD card at each sensor node before being transmitted to the gateway. We use SD cards in order to offer a large data storage capacity. Because SD modules are very high power-consuming, we combine the SD storage with a short-term 32KB EEPROM module. EEPROMs have the advantage of low power consumption but this comes at the cost of storage capacity. In order to cope with that, we use the EEPROM module as a temporary buffer to reduce the frequency of data storage on SD cards and hence reduce their power consumption.

LoRa communication

We use LoRa communication modules that operate on the 868Mhz frequency band and allow us to set the transmission power of the LoRa radio up to 23 dbm (200mW). This allows us to adapt the communication range of nodes and their power consumption to the deployment case. Due to the limited computation resources on our nodes, our first prototype supports only static routing between sensor nodes. It is noteworthy that when using LoRa communications, users are limited in terms of time usage of the frequency band because of regulations [116]. Due to this duty cycle constraint, we design the first version of the network in such a way that communication is unidirectional from sensor nodes to the gateway.

Sensing modules

We use in the first prototype of our platform 3 different sensing modules: an NO_2 analog sensing probe from Alphasense, an NO_2 digital sensing module from Cairpol and the DHT22 temperature and humidity digital sensing probe. The Alphasense sensor is connected to the ADC of the core module of measurement nodes in order to extract its

output voltage which is by default in the range 0-5V. In order to reduce the sensitivity of the Alphasense sensor to the variations of the input voltage of the circuit, we use decoupling capacitors. As for the Cairpol sensor, it communicates NO_2 concentrations in the range 0-250ppb using the UART protocol. We therefore develop a software routine to extract its data based on the specification of this communication protocol.

Power module

Our measurement nodes are powered using both a solar panel and batteries. Batteries are used to ensure that sensors work properly for at least 1 or 2 months before the batteries get drained. Solar panels are indeed used to extend the operation of the nodes especially during summer. In order to switch efficiently between the two power sources, we use an electronic circuit which allows to charge the battery using the solar energy during daytime and switch to battery power mode during nighttime.

3.2.4 Deployment case study and first results and feedbacks

We deployed 12 measurement nodes and one gateway in the Garibaldi street in Lyon from mid-July to the beginning of December 2018. The main sensors' circuit of the deployed prototype and the locations of the nodes are depicted in Fig. 3.5. The number of the deployed nodes was limited by our project budget and we chose to perform the deployment in the Garibaldi street because it is one of the main urban streets in Lyon in addition to being surrounded by two reference stations of the Lyon city. The latter allows us to compare our measurements to reference ones. Our streetwise deployment allows us to tackle our objective regarding the comparison between sensor measurements and physical models' simulations. Indeed, physical models usually provide the same concentration values for locations belonging to the same street [117]. Therefore, the 12 point measurements of our platform can help our fluid mechanics partners calibrate their simulation models.

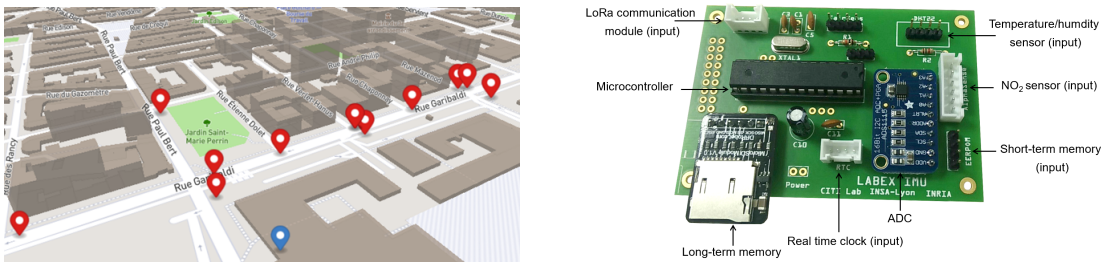


Figure 3.5: Deployment case study of the UrPolSens platform. From left to right: (a) the map of the 12 sensors (red) and the gateway (blue) that we deployed in the downtown of the Lyon city; and (b) a zoom on the main circuit of our sensor nodes.

Since our first deployment was limited to 12 sensor nodes within a small region, and in order to reduce the power consumption due to LoRa communications, we used TDMA as a channel access method. Sensors activate their radio only to send their data once every 10 minutes using a transmission power equal to 23dbm. Each sensor packet contains 10 average measurements of the last 10 minutes. With this communication

configuration, we allowed the nodes to operate on batteries for around 2 months. After that, the nodes continued their measurements while powered using solar panels.

In the first analysis of the data that we collected using our platform, we try to validate the nitrogen dioxide measurements that are provided by the energy efficient and low-cost sensing probes of Alphasense. To that end, we design an algorithm which converts raw data which is in the range of 0-5V to pollution concentrations in ppb. Our conversion takes into account the high impact of temperature on sensor measurements. Indeed, we noticed in our dataset that when temperature increases, the output voltage of the Alphasense sensors reacts poorly to the variations of nitrogen dioxide. Based on this fact, we propose to make a drift to the output voltage depending on the ambient temperature. We compare the corrected Alphasense data to the medium-level quality measurements of the Cairpol sensors and we depict the results in Fig. 3.6. We can clearly see that the Cairpol measurements and Alphasense calibrated data present a good correlation.

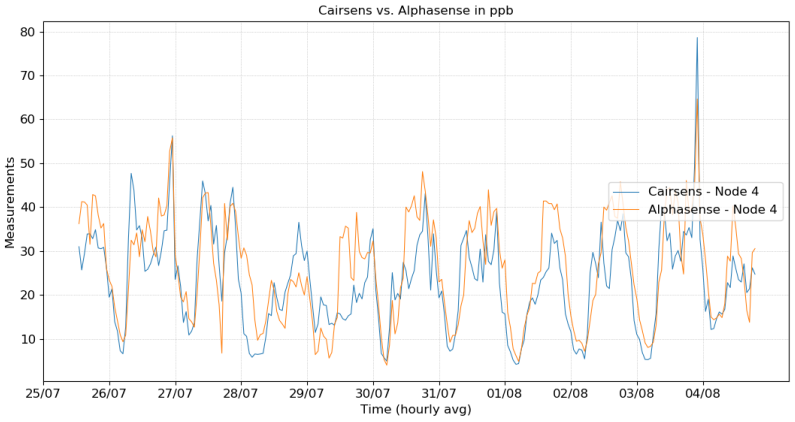


Figure 3.6: Comparison between Alphasense and Cairpol measurements.

In addition to the impact of environmental conditions on both sensing quality and transmission links, we also noticed some synchronization problems between sensor nodes and the gateway. Those problems happened in fact because of the real time clock of some nodes which was drifting too much in short time periods going to around 20s per day. After reviewing the datasheet of the RTC component, we understood that the reason behind this was the high temperatures during the month of August. Indeed, the temperatures were very severe because the nodes were exposed to the sun most of the day in order to maximize the solar energy usage.

3.3 Conclusion

In this chapter, we provided a summary of the main aspects of a selection of the research projects which focus on the design of WSN-based air pollution monitoring systems. Moreover, we presented our monitoring platform while highlighting our motivations compared to existing ones in terms of financial cost, energy consumption and modularity. We showed the architecture of our platform and explained how our design makes our system cost-effective, energy-efficient and modular. Finally, we performed

a deployment case study in the city of Lyon and showed that the first obtained results with our low-cost design are quite accurate while maintaining the energy efficiency of the system. Although in the first prototype the nodes were static, our sensors can be also mounted on mobile vehicles and the platform can be further extended to be used in a crowdsensing context.

The design of the UrPolSens platform allowed us to understand the specific characteristics of air pollution sensors mainly the operation and the energy consumption of the sensing probes. In the following chapters, we leverage this knowledge in order to design and evaluate application-aware WSN deployment and scheduling approaches.

Chapter 4

Deployment of WSN for Air Quality Mapping

In this chapter, we focus on the application of air quality mapping where the objective of the monitoring is to construct real-time air pollution maps by interpolating pollution sensor measurements. We tackle the WSN deployment problem and propose efficient optimization models and algorithms. Unlike most of the existing deployment approaches that we reviewed in chapter 2, our deployment solution is based on pollution-aware interpolation, which allows us to define an adequate mathematical formulation of pollution coverage quality.

We formulate the quality of air pollution mapping of a given sensor network depending on the sensing error of the deployed nodes and the estimation error of pollution concentration at locations where no sensor is deployed. The estimation error is defined as the difference between the ground truth (or real) value of pollution concentration and the concentration obtained by applying an adequate interpolation method on the measurements of sensor nodes. We derive from our formulation of coverage quality two deployment models using integer programming modeling. The first deployment model allows us to minimize the network deployment cost while ensuring a required coverage quality. In the second model, we propose instead to optimize the coverage quality without exceeding a given deployment budget. Both optimization models ensure the connectivity of the network based on the flow concept, which guarantees that the deployed sensors can send pollution data to at least one sink node. We also take into account the sensing error of sensor nodes, the impact of weather conditions on the variations of pollution concentrations and the heterogeneity of sensing and communication characteristics of nodes.

We analyze the theoretical complexity of our models and propose heuristic algorithms based on linear programming relaxation and the concept of binary search. We perform extensive simulations on a dataset of the Lyon city in order to assess the computational complexity of the proposed models and algorithms. We also evaluate the performance of our optimization approach in terms of coverage and connectivity in order to derive engineering insights on the optimal WSN deployment.

4.1 Mathematical formulation of air pollution mapping quality

Most of the existing deployment approaches are generic and do not take into account the nature of the phenomenon to be covered. In order to cope with that and define an appropriate formulation of the coverage quality of a given sensor network topology in the context of air pollution mapping, we propose in this section to use interpolation. We formulate the quality of air pollution mapping based on the sensing error of the deployed nodes and the estimation error of pollution concentration at locations where no sensor is deployed.

4.1.1 Characterization of the deployment region

We consider as input the map of a given urban area that we call the deployment region. Let $\mathcal{P}$ be a set of discrete points approximating the deployment region at a high-scale ($|\mathcal{P}| = \mathcal{N}$). The set $\mathcal{P}$ can be obtained using a 2D or 3D discretization.

Our goal in this chapter is to be able to determine with a high precision the concentration value at each point $p \in P$. We ensure that for each point $p \in P$, either a sensor is deployed, or the pollution concentration can be estimated with a high precision based on the data gathered by the neighboring deployed sensors. We also ensure that all the deployed sensors can send their data to at least one sink node while optimizing the positions of sink nodes.

In general case, the set $\mathcal{P}$ is considered as the set of potential positions of WSN nodes. However, in smart cities applications, some restrictions on node positions may apply because of authorization or practical issues. For instance, in order to alleviate the energy constraints, we may place sensors on only lampposts and traffic lights as experimented in [118]. When this is the case, we do not consider as potential positions the points $p \in P$ where sensors cannot be deployed.

We use decision variables x_p (respectively y_p) to specify if a sensor (respectively a sink) is deployed at point p or not. The main notations used in this section are presented in TABLE 4.1.

4.1.2 Interpolation formulation

Our idea is to base on interpolation in order to formulate coverage quality based on the sensing error of the deployed nodes and the estimation error of pollution concentration at locations where no sensor is deployed. Interpolation formulate the estimated concentration $\widehat{\mathcal{Z}}_p$ at a given location $p \in \mathcal{P}$ as a weighted combination of the measured concentrations $\mathcal{Z}_q$, $q \in \mathcal{P}$ [119]. The weights of the measured concentrations $\mathcal{W}_{pq}$ are called correlation coefficients and can be evaluated in a deterministic way based on the distance between the location of the measured concentration and the location of the estimated concentration. In this case, which is called the Inverse Distance Weighting interpolation, $\widehat{\mathcal{Z}}_p$ is evaluated using formula 4.1. The correlation coefficients can be also evaluated in a stochastic way, the most used method doing so is called kriging. Without loss of generality, we focus in this chapter on the case of deterministic interpolation.

Sets and parameters	
$\mathcal{P}$	Set of points (the deployment region)
$\mathcal{N}$	Number of points
$\mathcal{K}$	Set of sensor types
$\mathcal{T}$	Set of time snapshots
$\mathcal{G}_p$	Ground truth pollution concentrations
$\mathcal{M}_p^t$	Simulated pollution concentrations (using atmospheric dispersion simulation)
$\mathcal{Z}_p$	Measured pollution concentrations
$\widehat{\mathcal{Z}}_p$	Estimated pollution concentrations (using interpolation)
m_p^t	Simulation errors (using atmospheric dispersion simulation)
s_p^k	Sensing errors
$\mathcal{W}_{pq}$	Correlation coefficients
$\mathcal{B}_{pq}$	Correlation-existence coefficients
$\mathcal{D}$	The correlation distance function
d	Maximum correlation distance
α	Attenuation coefficient of the correlation distance
β	Wind coefficient of correlation
$\Gamma^k(p)$	Communication neighborhoods
$\mathcal{R}^k$	Communication ranges
A_{pq}^k	Adjacency matrix
E	Tolerated estimation error
I	The maximum number of sinks
δ_p^k	The cost of sensors
ψ_p	The cost of sinks
$\mathcal{F}$	Deployment cost of the whole network
J	Deployment budget
Decision variables of optimization models	
x_p^k	Define whether a sensor of type k is deployed at point p or not ; $x_p^k \in \{0,1\},\ p \in \mathcal{P}$
y_p	Define whether a sink is deployed at point p or not; $y_p \in \{0,1\},\ p \in \mathcal{P}$
Auxiliary variables of optimization models	
e_p	Estimation error of pollution concentration at point p; $e_p \geq 0,\ p \in \mathcal{P}$
g_{pq}	Flow quantity transmitted from node p to node q $g_{pq} \in \{0,1,...\},\ p \in \mathcal{P},\ q \in \Gamma(p)$

Table 4.1: Main notations used in the air quality mapping approach.

$$\widehat{\mathcal{Z}}_p = \frac{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} * \mathcal{Z}_q}{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq}} \tag{4.1}$$

Formula 4.1 can be generalized in order to take into account that: 1) $\widehat{\mathcal{Z}}_p = \mathcal{Z}_p$ when $x_p = 1$; and 2) $\mathcal{Z}_q$ values should be considered in the fraction only when $x_q = 1$. Hence, we get formula 4.2 in addition to constraint 4.3 that ensures that the denominator is never equal to zero. $\mathcal{B}_{pq}$ parameters define whether there is a correlation between points p and q or not; that is, $\mathcal{B}_{pq} = 1$ when $\mathcal{W}_{pq} > 0$.

$$\widehat{\mathcal{Z}}_p = x_p \cdot \mathcal{Z}_p + (1 - x_p) \cdot \frac{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot x_q \cdot \mathcal{Z}_q}{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot x_q} \tag{4.2}$$

$$\sum_{q \in \mathcal{P}} \mathcal{B}_{pq} \cdot x_q \geq 1 \tag{4.3}$$

4.1.3 Correlation coefficients

In order to take into account the context of air pollution estimation, we propose to calculate the $\mathcal{W}_{pq}$ parameters using formula 4.4 based on the wind direction and the distance between the two points p and q. $\mathcal{D}(p, q)$ is the distance function. α is the attenuation coefficient of the correlation distance, this means that for greater values of α, very low correlation coefficients are assigned to far points. $\mathcal{B}_{pq}$ is calculated based on d, which is called the maximum correlation distance and defines the range of the correlated neighboring points of a given point. The last parameter of formula 4.4 is β, which allows us to take into account the direction of wind. As in [120] and for simplicity's sake, we set $\beta = 1$ if the wind direction is $\vec{pq}$ or $\vec{qp}$, otherwise $\beta = 0$.

$$\mathcal{W}_{pq} = \mathcal{B}_{pq} * (\frac{1}{\mathcal{D}(p, q)^{\alpha}} + \beta) \quad \text{if } p \neq q \tag{4.4}$$

$$\mathcal{B}_{pq} = \begin{cases} 1 \text{ if } q \in Disc(p, d) \\ 0 \text{ if } q \notin Disc(p, d) \end{cases} \tag{4.5}$$

In order to take into account the impact of the urban topography on the dispersion of pollutants, $\mathcal{D}$ can be defined as the shortest distance along the roads network. This metric was used in [121] to assign small correlation values to points that are separated by buildings, even if they are close with respect to the Euclidean distance.

4.1.4 Ground truth, simulated and measured pollution concentrations

Let $\mathcal{G}_p$ denote the unknown ground truth (or real) value of pollution concentration at point p and $\mathcal{M}_p$ denote the simulated pollution concentration obtained by means of an atmospheric dispersion simulator. The relationship between $\mathcal{G}_p$ and $\mathcal{M}_p$ (respectively $\mathcal{G}_p$ and $\mathcal{Z}_p$) is presented in formula 4.6 (respectively in formula 4.7) where m_p is the maximum simulation error given by the dispersion simulator at point p and s_p is the maximum sensing error in the measurements of a node deployed at point p.

$$\mathcal{M}_p - \mathcal{G}_p \in [-m_p, m_p] \tag{4.6}$$
$$\mathcal{Z}_p - \mathcal{G}_p \in [-s_p, s_p] \tag{4.7}$$

4.1.5 Basic coverage quality formulation

We define the quality of pollution coverage at point p as the maximum estimation error e_p of the interpolation method; i.e. the absolute difference between the estimated concentration $\widehat{\mathcal{Z}}_p$ and the ground truth concentration $\mathcal{G}_p$. Hence, we get formula 4.8 where UB is the least upper bound (or maximum) function.

$$e_p = UB(|\widehat{\mathcal{Z}}_p - \mathcal{G}_p|) \tag{4.8}$$

In order to be able to use this coverage quality measure in the optimization models, we need to express it depending only on the known parameters, this means that we should not use parameters $\mathcal{Z}$ and $\mathcal{G}$ in the estimation error formulation. First we simplify formula 4.8 to obtain formula 4.9 and then formula 4.10 based on the fact that either $x_p = 1$ or $1 - x_p = 1$. Then, we get formula 4.11 based on the relationship between $\mathcal{Z}_p$ and $\mathcal{G}_p$ presented in formula 4.7.

$$\begin{aligned} e_p = \; & UB(|x_p \cdot (\mathcal{Z}_p - \mathcal{G}_p) \\ & +(1 - x_p) \cdot \frac{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot x_q \cdot (\mathcal{Z}_q - \mathcal{G}_p)}{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot x_q}|) \end{aligned} \tag{4.9}$$

$$\begin{aligned} e_p = \; & x_p \cdot UB(|\mathcal{Z}_p - \mathcal{G}_p|) \\ & +(1 - x_p) \cdot \frac{UB(|\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot x_q \cdot (\mathcal{Z}_q - \mathcal{G}_p)|)}{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot x_q} \end{aligned} \tag{4.10}$$

$$e_p = x_p \cdot s_p + (1 - x_p) \cdot \frac{UB(|\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot x_q \cdot (\mathcal{Z}_q - \mathcal{G}_p)|)}{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot x_q} \tag{4.11}$$

In order to eliminate $\mathcal{Z}_q$ and $\mathcal{G}_p$ from 4.11, notice that the numerator of the fraction can be simplified to get formula 4.12. Also, notice that the members of the summations in formula 4.12 can be simplified as shown in formulas 4.13 and 4.14. At the end, the expression of the estimation error depending only on known parameters is given in formula 4.15.

$$
\begin{aligned}
&UB(|\sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q\cdot(\mathcal{Z}_q-\mathcal{G}_p)|)\\
&= Max\ \{UB(\sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q\cdot(\mathcal{Z}_q-\mathcal{G}_p)),\\
&\quad UB(-\sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q\cdot(\mathcal{Z}_q-\mathcal{G}_p))\}\\
&= Max\{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q\cdot UB(\mathcal{Z}_q-\mathcal{G}_p),\\
&\quad \sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q\cdot UB(\mathcal{G}_p-\mathcal{Z}_q)\}
\end{aligned}
\tag{4.12}
$$

$$
\begin{aligned}
UB(\mathcal{Z}_q-\mathcal{G}_p) &= UB(\mathcal{M}_q-\mathcal{M}_p+\mathcal{Z}_q-\mathcal{M}_q+\mathcal{M}_p-\mathcal{G}_p)\\
&= \mathcal{M}_q-\mathcal{M}_p+UB(\mathcal{Z}_q-\mathcal{M}_q)+UB(\mathcal{M}_p-\mathcal{G}_p)\\
&= \mathcal{M}_q-\mathcal{M}_p+s_q+m_q+m_p
\end{aligned}
\tag{4.13}
$$

$$
UB(\mathcal{G}_p-\mathcal{Z}_q) \;=\; \mathcal{M}_p-\mathcal{M}_q+s_q+m_q+m_p \tag{4.14}
$$

$$
\begin{aligned}
e_p = x_p\cdot s_p + \frac{(1-x_p)}{\sum_{q\in\mathcal{P}}\mathcal{W}_{pq}\cdot x_q}\cdot\ Max\ \{&\\
&\sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q\cdot(\mathcal{M}_q-\mathcal{M}_p+s_q+m_q+m_p),\\
&\sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q\cdot(\mathcal{M}_p-\mathcal{M}_q+s_q+m_q+m_p)\}
\end{aligned}
\tag{4.15}
$$

4.1.6 Multi-scenario coverage quality formulation

In the expression of the quality of pollution coverage at point p given in formula 4.15, the values of $\mathcal{M}_p$ and m_p are considered as constants. We recall that $\mathcal{M}_p$ and m_p are obtained by simulation depending on pollution emissions and weather conditions. This means that when the sensor network is operating at a given moment t, the estimation error e_p corresponds to the expression presented in 4.15 only if $\mathcal{M}_p$ and m_p values are obtained while considering the weather conditions and pollution emissions that correspond to the moment t. In order to cope with that, we first propose to consider a set of time snapshots $\mathcal{T}$ where each snapshot $t \in \mathcal{T}$ corresponds to a potential scenario of

weather conditions and pollution emissions. Then, we define $\mathcal{M}_p^t$ and m_p^t as, respectively, the simulated pollution concentration and the maximum simulation error at point p while considering weather conditions and pollution emissions corresponding to the snapshot t. Finally, we define in formula 4.16 the new formulation of pollution coverage quality as the maximum error among all the time snapshots $\mathcal{T}$. Note that the number of the time snapshots is a key factor in this new definition of coverage quality. Indeed, the more the number of snapshots, the better is the approximation of the variables $\mathcal{M}_p$ and m_p.

$$
\begin{aligned}
e_p = \; & Max_{\,t\in\mathcal{T}}\{ \\
& x_p \cdot s_p + \frac{(1-x_p)}{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q} \cdot \; Max\,\{ \\
& \quad \sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q \cdot (\mathcal{M}_q^t - \mathcal{M}_p^t + s_q + m_q^t + m_p^t), \\
& \quad \sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot x_q \cdot (\mathcal{M}_p^t - \mathcal{M}_q^t + s_q + m_q^t + m_p^t)\}\}
\end{aligned}
\tag{4.16}
$$

4.1.7 Taking into account sensing heterogeneity

So far, we have been considering that all the sensor nodes have the same sensing error at a given point p. The coverage quality formulation can be more general by considering a set of sensor types $\mathcal{K}$ where the sensing error depends on the type of the sensor; i.e the sensing error of a sensor of type $k \in \mathcal{K}$ at point $p \in \mathcal{P}$ is denoted s_p^k. In this case, the index k is also added to variables x_p in order to denote the fact that a sensor of type k is placed at point p, hence we get x_p^k variables. Note that only one type can be chosen to be deployed at point p as formulated in formula 4.17.

$$
\sum_{k\in\mathcal{K}} x_p^k \leq 1, \qquad p \in \mathcal{P} \tag{4.17}
$$

In order to take into account the heterogeneity of sensor nodes in the formulation of coverage quality, we transform formulas 4.16 and 4.3 to formulas 4.18 and 4.19 by adding the index k to the sensing errors and replacing x_p variables by $\sum_{k\in\mathcal{K}} x_p^k$.

$$
\begin{aligned}
e_p = \; & Max_{\,t\in\mathcal{T}}\{ \\
& \sum_{k\in\mathcal{K}} x_p^k \cdot s_p^k + \frac{(1-\sum_{k\in\mathcal{K}} x_p^k)}{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot \sum_{k\in\mathcal{K}} x_q^k} \cdot \; Max\,\{ \\
& \quad \sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot \sum_{k\in\mathcal{K}} x_q^k \cdot (\mathcal{M}_q^t - \mathcal{M}_p^t + s_q^k + m_q^t + m_p^t), \\
& \quad \sum_{q\in\mathcal{P}} \mathcal{W}_{pq}\cdot \sum_{k\in\mathcal{K}} x_q^k \cdot (\mathcal{M}_p^t - \mathcal{M}_q^t + s_q^k + m_q^t + m_p^t)\}\}
\end{aligned}
\tag{4.18}
$$

$$\sum_{q \in \mathcal{P}} \mathcal{B}_{pq} \cdot \sum_{k \in \mathcal{K}} x_q^k \geq 1, \qquad p \in \mathcal{P} \tag{4.19}$$

4.2 Optimization models

In this section, we use integer programming modeling to derive two optimization models for WSN deployment based on the formulation of estimation errors e_p that we presented in the previous section in formula 4.18. The first deployment model allows us to minimize the network deployment cost while ensuring a required coverage quality. In the second model, we propose instead to optimize the coverage quality without exceeding a given deployment budget. Both optimization models ensure the connectivity of the network using the flow concept while taking into account the heterogeneous communication capabilities of nodes.

4.2.1 MIN_COST: Deployment cost minimization

Objective function

We first denote by δ_p^k (respectively ψ_p) the deployment cost of a sensor of type k (respectively a sink) at point p. The network deployment cost to minimize is thus given as follows:

$$\mathcal{F} = \sum_{p \in \mathcal{P}} \sum_{k \in \mathcal{K}} \delta_p^k * x_p^k + \sum_{p \in \mathcal{P}} \psi_p * y_p \tag{4.20}$$

Air pollution mapping constraints

First, let the function $f(\{e_p,\ p \in \mathcal{P}\})$ denote the estimation error of a given sensor network topology. f is calculated based on the estimation error at all the points that define the deployment region. Without loss of generality, we consider in this chapter that f is either the maximum or the mean function. The main constraint of air pollution mapping is defined in formula 4.21 and ensures that the estimation error in the deployment region does not exceed the required precision E that we call the tolerated estimation error. The function f is defined using constraint 4.22 (respectively 4.23) in the case of the max function (respectively the mean function) where $\mathcal{N} = |\mathcal{P}|$.

$$f \leq E \tag{4.21}$$

$$f \geq e_p, \quad p \in \mathcal{P} \tag{4.22}$$

$$f = \textstyle\sum_{p \in \mathcal{P}} e_p / \mathcal{N} \tag{4.23}$$

Constraint 4.21 together with constraint 4.18 defined in the previous section ensure that the pollution estimation error of the resulting network is bounded by the tolerated error E. However, constraint 4.18 is too complex and should be linearized in order to

be solved efficiently by a mathematical solver and this by eliminating the Max and division functions in addition to the product between variables x_p^k, x_q^k and e_p. Note that since we are constraining e_p variables to minimum values, the equality in constraint 4.18 can be transformed to an inequality as in constraint 4.24. Therefore, the Max function can be easily eliminated and thus we get constraints 4.25 and 4.26.

$$
\begin{aligned}
& e_p \geq \ Max\ _{t\in\mathcal{T}}\{ \\
& \sum_{k\in\mathcal{K}} x_p^k \cdot s_p^k + \frac{(1-\sum_{k\in\mathcal{K}} x_p^k)}{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot \sum_{k\in\mathcal{K}} x_q^k} \cdot \ Max\ \{ \\
& \quad \sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot \sum_{k\in\mathcal{K}} x_q^k \cdot (\mathcal{M}_q^t - \mathcal{M}_p^t + s_q^k + m_q^t + m_p^t), \\
& \quad \sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot \sum_{k\in\mathcal{K}} x_q^k \cdot (\mathcal{M}_p^t - \mathcal{M}_q^t + s_q^k + m_q^t + m_p^t)\} \\
& \}
\end{aligned}
\tag{4.24}
$$

$$
\begin{aligned}
& e_p \geq \sum_{k\in\mathcal{K}} x_p^k \cdot s_p^k + \frac{(1-\sum_{k\in\mathcal{K}} x_p^k)}{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot \sum_{k\in\mathcal{K}} x_q^k} \cdot \\
& \quad \sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot \sum_{k\in\mathcal{K}} x_q^k \cdot (\mathcal{M}_q^t - \mathcal{M}_p^t + s_q^k + m_q^t + m_p^t) \\
& , \ p \in \mathcal{P},\ t \in \mathcal{T}
\end{aligned}
\tag{4.25}
$$

$$
\begin{aligned}
& e_p \geq \sum_{k\in\mathcal{K}} x_p^k \cdot s_p^k + \frac{(1-\sum_{k\in\mathcal{K}} x_p^k)}{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot \sum_{k\in\mathcal{K}} x_q^k} \cdot \\
& \quad \sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot \sum_{k\in\mathcal{K}} x_q^k \cdot (\mathcal{M}_p^t - \mathcal{M}_q^t + s_q^k + m_q^t + m_p^t) \\
& , \ p \in \mathcal{P},\ t \in \mathcal{T}
\end{aligned}
\tag{4.26}
$$

We simplify constraints 4.25 and 4.26 by multiplying their both sides by the denominator; thus we get, respectively, constraints 4.27 and 4.28.

$$
\begin{aligned}
& \sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot \sum_{k\in\mathcal{K}} x_q^k \cdot (e_p - \sum_{k'\in\mathcal{K}} x_p^{k'} \cdot s_p^{k'} - \\
& (1 - \sum_{k'\in\mathcal{K}} x_p^{k'}) \cdot (\mathcal{M}_q^t - \mathcal{M}_p^t + s_q^k + m_q^t + m_p^t)) \geq 0 \\
& , \ p \in \mathcal{P},\ t \in \mathcal{T}
\end{aligned}
\tag{4.27}
$$

$$
\begin{aligned}
& \sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot \sum_{k\in\mathcal{K}} x_q^k \cdot (e_p - \sum_{k'\in\mathcal{K}} x_p^{k'} \cdot s_p^{k'} - \\
& (1 - \sum_{k'\in\mathcal{K}} x_p^{k'}) \cdot (\mathcal{M}_p^t - \mathcal{M}_q^t + s_q^k + m_q^t + m_p^t)) \geq 0 \\
& , \ p \in \mathcal{P},\ t \in \mathcal{T}
\end{aligned}
\tag{4.28}
$$

Now, we have to eliminate the product between variables in order to get a linear formulation. To that end, we first define h_{pq}^{kt} and l_{pq}^{kt} variables in formulas 4.29 and 4.30. Based on the definition of the latter variables, we transform constraints 4.27 and 4.28 to constraints 4.31 and 4.32.

$$\begin{aligned} h_{pq}^{kt} = x_q^k \cdot (e_p - \sum_{k' \in \mathcal{K}} x_p^{k'} \cdot s_p^{k'} - \\ (1 - \sum_{k' \in \mathcal{K}} x_p^{k'}) \cdot (\mathcal{M}_q^t - \mathcal{M}_p^t + s_q^k + m_q^t + m_p^t)) \\ , \ (p,q) \in \mathcal{P} : \mathcal{B}pq = 1, \ k \in \mathcal{K}, \ t \in \mathcal{T} \end{aligned} \tag{4.29}$$

$$\begin{aligned} l_{pq}^{kt} = x_q^k \cdot (e_p - \sum_{k' \in \mathcal{K}} x_p^{k'} \cdot s_p^{k'} - \\ (1 - \sum_{k' \in \mathcal{K}} x_p^{k'}) \cdot (\mathcal{M}_p^t - \mathcal{M}_q^t + s_q^k + m_q^t + m_p^t)) \\ , \ (p,q) \in \mathcal{P} : \mathcal{B}pq = 1, \ k \in \mathcal{K}, \ t \in \mathcal{T} \end{aligned} \tag{4.30}$$

$$\sum_{q \in \mathcal{P} : \mathcal{B}_{pq} = 1} \mathcal{W}_{pq} \cdot \sum_{k \in \mathcal{K}} h_{pq}^{kt} \geq 0, \qquad p \in \mathcal{P}, \ t \in \mathcal{T} \tag{4.31}$$

$$\sum_{q \in \mathcal{P} : \mathcal{B}_{pq} = 1} \mathcal{W}_{pq} \cdot \sum_{k \in \mathcal{K}} l_{pq}^{kt} \geq 0, \qquad p \in \mathcal{P}, \ t \in \mathcal{T} \tag{4.32}$$

We now linearize the definition of variables h_{pq}^{kt} in constraints 4.33 and 4.34 where H is a big number defined to relax constraint 4.33 when $x_q^k = 1$ and constraint 4.34 when $x_q^k = 0$.

$$-H * x_q^k \leq h_{pq}^{kt} \leq H * x_q^k, \ (p,q) \in \mathcal{P} : \mathcal{B}_{pq} = 1, \ k \in \mathcal{K}, \ t \in \mathcal{T} \tag{4.33}$$

$$\begin{aligned} -H * (1 - x_q^k) + (e_p - \textstyle\sum_{k' \in \mathcal{K}} x_p^{k'} \cdot s_p^{k'} - \\ (1 - \textstyle\sum_{k' \in \mathcal{K}} x_p^{k'}) \cdot (\mathcal{M}_q^t - \mathcal{M}_p^t + s_q^k + m_q^t + m_p^t)) \\ \leq h_{pq}^{kt} \leq \\ H * (1 - x_q^k) + (e_p - \textstyle\sum_{k' \in \mathcal{K}} x_p^{k'} \cdot s_p^{k'} - \\ (1 - \textstyle\sum_{k' \in \mathcal{K}} x_p^{k'}) \cdot (\mathcal{M}_q^t - \mathcal{M}_p^t + s_q^k + m_q^t + m_p^t)), \\ (p,q) \in \mathcal{P} : \mathcal{B}_{pq} = 1, \ k \in \mathcal{K}, \ t \in \mathcal{T} \end{aligned} \tag{4.34}$$

The same thing is performed on l_{pq}^{kt} variables to get constraints 4.35 and 4.36 where L is a big number defined to relax constraint 4.35 when $x_q^k = 1$ and constraint 4.36

when $x_q^k = 0$. At the end, the linear formulation of air pollution mapping is ensured by constraints 4.17, 4.19, 4.21, 4.31, 4.32, 4.33, 4.34, 4.35 and 4.36.

$$-L * x_q^k \leq l_{pq}^{kt} \leq L * x_q^k,\ (p,q) \in \mathcal{P} : \mathcal{B}_{pq} = 1,\ k \in \mathcal{K},\ t \in \mathcal{T} \tag{4.35}$$

$$\begin{gathered} -L * (1 - x_q^k) + (e_p - \textstyle\sum_{k' \in \mathcal{K}} x_p^{k'} \cdot s_p^{k'} - \\ (1 - \textstyle\sum_{k' \in \mathcal{K}} x_p^{k'}) \cdot (\mathcal{M}_p^t - \mathcal{M}_q^t + s_q^k + m_q^t + m_p^t)) \\ \leq l_{pq}^{kt} \leq \\ L * (1 - x_q^k) + (e_p - \textstyle\sum_{k' \in \mathcal{K}} x_p^{k'} \cdot s_p^{k'} - \\ (1 - \textstyle\sum_{k' \in \mathcal{K}} x_p^{k'}) \cdot (\mathcal{M}_p^t - \mathcal{M}_q^t + s_q^k + m_q^t + m_p^t)), \\ (p,q) \in \mathcal{P} : \mathcal{B}_{pq} = 1,\ k \in \mathcal{K},\ t \in \mathcal{T} \end{gathered} \tag{4.36}$$

Network connectivity constraints

We formulate the connectivity constraint as a network flow problem. We consider the same potential positions set $\mathcal{P}$ for sensors and sinks. We first denote by $\Gamma^k(p)$ where $p \in \mathcal{P}$ and $k \in \mathcal{K}$, the set of neighbors of a sensor node of type k deployed at point p. This set can be determined using sophisticated propagation models. It can be also determined using the binary disc model, in which case $\Gamma^k(p) = \{q \in P \text{ where } q \in Disc(p, R^k)\}$ where R^k is the communication range of a sensor of type k. Let A_{pq}^k be equal to 1 if $q \in \Gamma^k(p)$ and equal to 0 otherwise. Now, we define the flow variables g_{pq} as the flow quantity transmitted from a node located at point p to another node located at point q. The idea is to suppose that each sensor of the resulting WSN generates a flow unit in the network, and then verify if these units can be recovered by sink nodes. The following constraints ensure that the deployed sensors and sinks form a connected wireless sensor network; i.e. each sensor can communicate with at least one sink.

$$g_{pq} \leq N * (y_p + \textstyle\sum_{k \in \mathcal{K}} x_p^k * A_{pq}^k),\ (p,q) \in \widehat{\mathcal{P}} \tag{4.37}$$

$$g_{pq} \leq N * (y_q + \textstyle\sum_{k \in \mathcal{K}} x_q^k * A_{qp}^k),\ (p,q) \in \widehat{\mathcal{P}} \tag{4.38}$$

$$\begin{gathered} \textstyle\sum_{k \in \mathcal{K}} x_p^k - \mathcal{N} * y_p \\ \leq \textstyle\sum_{q \in \widehat{\mathcal{P}}} g_{pq} - \sum_{q \in \widehat{\mathcal{P}}} g_{qp} \leq \\ \textstyle\sum_{k \in \mathcal{K}} x_p^k,\ p \in \widehat{\mathcal{P}} \end{gathered} \tag{4.39}$$

$$\textstyle\sum_{p \in \widehat{\mathcal{P}}} \sum_{q \in \widehat{\mathcal{P}}} g_{pq} = \sum_{p \in \widehat{\mathcal{P}}} \sum_{q \in \widehat{\mathcal{P}}} g_{qp} \tag{4.40}$$

$$\textstyle\sum_{p \in \widehat{\mathcal{P}}} y_p \leq I \tag{4.41}$$

Constraint 4.37 (respectively constraint 4.38) forces to 0 the flow from p to q if neither a sensor nor a sink is positioned at point p (respectively at point q) or a sensor of

type k is positioned at point p (respectively at point q) but cannot communicate with its neighboring point q (respectively point p). The assumption of sensors that should generate, each of which, a flow unit is guaranteed thanks to constraint 4.39, which is relaxed if a sink is deployed at point p; that is a sink can receive up to N flow units, which is the maximum number of sensors that we can have in the network. The overall flow is conservative thanks to constraint 4.40, this means that sinks recover all the flow units that are generated by sensor nodes. Finally, constraint 4.41 allows us to fix the maximum number of sinks of the resulting network that we denote I. This constraint can be relaxed by setting I to N, in which case the mixed integer programming model optimizes the number of sink nodes in addition to the optimization of their positions.

MILP model

Finally, the mixed integer programming model that allows us to minimize the network deployment cost while ensuring a required coverage quality is denoted MIN_COST and can be written as follows:

[MIN_COST]
Objective: minimize $\mathcal{F}$
Pollution mapping constraints: $(4.17), (4.19), (4.21), (\text{cf})^*,$
$(4.31), (4.32), (4.33), (4.34), (4.35), (4.36)$
Connectivity constraints: $(4.37), (4.38), (4.39), (4.40), (4.41)$
Decision variables: $x_p^k, y_p \in \{0, 1\}$
Auxiliary variables: $e_p, g_{pq} \in \mathbb{R}^+; h_{pq}^{kt}, l_{pq}^{kt} \in \mathbb{R}$
* : as explained in the beginning of section 4.2.1,
(cf) should be replaced by either (4.22) or (4.23)

4.2.2 MIN_ERROR: Estimation error minimization

In the second MILP model, we propose to minimize the estimation error in the deployment region f defined in formulas 4.22 and 4.23. The deployment cost of the network should be constrained by the budget J as in constraint 4.42.

$$\mathcal{F} \leq J \tag{4.42}$$

Therefore, the second MILP model, denoted MIN_ERROR, is written as follows:

[MIN_ERROR]
Objective: minimize f
Budget constraint: (4.42)
Pollution mapping constraints: (4.17), (4.19), (cf)*, (4.31), (4.32), (4.33), (4.34), (4.35), (4.36)
Connectivity constraints: (4.37), (4.38), (4.39), (4.40), (4.41)
Decision variables: $x_p^k, y_p \in \{0,1\}$
Auxiliary variables: $e_p, g_{pq} \in \mathbb{R}^+; h_{pq}^{kt}, l_{pq}^{kt} \in \mathbb{R}$
* : as explained in the beginning of section 4.2.1,
(cf) should be replaced by either (4.22) or (4.23)

4.3 Resolution of the optimization models

4.3.1 Exact MILP solvers and theoretical complexity

The proposed optimization models are based on integer linear programming that can be solved using the exact MILP solvers. In the performance evaluation section, we use the Cplex exact solver from IBM mainly to solve small instances of our deployment problem. Indeed, the MILP models are proved in the literature to be NP-hard, i.e. the execution time of the MILP solvers increases exponentially with the size of the problem. In our models, the size of the problem depends on the number of points approximating the deployment region $|\mathcal{P}|$, the number of time snapshots $|\mathcal{T}|$ and the number of sensor types $|\mathcal{K}|$. In fact, what makes the MILP models NP-hard is the number of binary variables which causes an exponential increase in the number of iterations when using the exact MILP solvers. Note that the two proposed models have the same size since MIN_ERROR is obtained by changing the objective function of MIN_COST and constraining the network deployment cost $\mathcal{F}$ instead of the estimation error f. In both of our models, the number of binary variables is equal to $|\mathcal{P}| \cdot |\mathcal{K}|$. That is, the complexity of the models is mainly due to the number of potential positions and the number of sensor types.

4.3.2 Linear-relaxation based heuristic

In order to solve our optimization models on large instances in a reasonable time while getting near-optimal solutions, we propose to use the concept of linear relaxation to design a resolution algorithm for the model MIN_COST. As for solving the model MIN_ERROR, we propose to use the concept of binary search in addition to linear relaxation. The linear relaxation consists of removing the integrality constraint on x_p and y_p variables, which allows our heuristics to run in polynomial time as we show later in this section.

Solving MIN_COST

We first define the linear programming model LP1 while considering the same objective function and constraints as MIN_COST and relaxing all the binary variables x_p^k and y_p; i.e. binary variables are considered real in the range of $[0, 1]$, this means that the solutions of the LP1 model are not necessarily binary. Note that in a given solution of LP1 where placement variables x_p^k and y_p are fractional, the variable having the maximum value (i.e. the closest binary variable to 1) corresponds to the most important node in the satisfaction of coverage and connectivity constraints. Based on this fact, we propose in each iteration of our heuristic algorithm presented in Algorithm 1 to set a sensor of type k_0 at point p_0 where $x_{p_0}^{k_0}$ is the closet variable to 1 or to set a sink at point p_0 if y_{p_0} is the closest variable to 1. The loop, which performs iterative rounding, stops once the placement variables are equal to either 0 or 1.

Algorithm 1 Heuristic algorithm to solve MIN_COST

Inputs: $\mathcal{P}, \mathcal{T}$
Outputs: $\{x_p^k\}, \{y_p\}$

repeat
  Solve the LP model
  Let v be the maximum fractional variable among x_p^k and y_p variables
  Add constraint $v = 1$ to the LP model
until all the variables are binary

The theoretical complexity of Algorithm 1 mainly depends on the number of iterations in the relaxation loop since solving the LP1 model by the exact solvers runs in polynomial time. Note that the number of iterations is at most equal to the number of points $\mathcal{P}$, which happens when a node has to be deployed at each point. As a result, Algorithm 1 runs in polynomial time.

Solving MIN_ERROR

We recall that MIN_ERROR is in fact the dual problem of MIN_COST where the objective function and the main coverage constraint have been exchanged. Let J be the deployment cost that we obtain by solving MIN_COST when the tolerated estimation error is set to E. In this case, E can be considered as a solution to the model MIN_ERROR when the deployment budget is set to J. We base on this duality and use the principle of binary search in order to define Algorithm 2 that allows us to solve the problem MIN_ERROR. The idea is to explore the interval of the possible values of the tolerated estimation error E which are in the range $[U, V]$. U and V define the bounds on the estimation error that we can get with the budget J. Note that the estimation error never exceeds $\max_{p\in\mathcal{P},t\in\mathcal{T}} \mathcal{M}_p^t + 2 * m_p^t$ according to our definition given in section 4.1. The interval $[U, V]$ is tightened in every iteration and the main loop stops once the length of the interval is sufficiently small compared to a given threshold TH.

The number of iterations is a key factor in the complexity of Algorithm 2 and is equal to $\gamma = [log_2(V_0/TH)]$ where $V_0 = \max_{p\in\mathcal{P},t\in\mathcal{T}} \mathcal{M}_p^t + 2 * m_p^t$. The value of the

Algorithm 2 Heuristic algorithm to solve MIN_ERROR

Inputs: $\mathcal{P}, \mathcal{T}$
Outputs: $\{x_p^k\}, \{y_p\}$

$U \leftarrow 0$
$V \leftarrow \max_{p \in \mathcal{P}, t \in \mathcal{T}} \mathcal{M}_p^t + 2 * m_p^t$
repeat
  $H \leftarrow \frac{U+V}{2}$
  if $OBJ\_MIN\_COST(H) \leq J$ **then**
    $V \leftarrow H$
  else
    $U \leftarrow H$
until $V - U \leq TH$
return $SOLUTION\_MIN\_COST(V)$

constant γ mainly depends on the nature of pollutants that defines the value of V_0. The theoretical complexity of Algorithm 2 is equal to γ multiplied by the complexity of solving MIN_COST that occurs in each iteration. As a result, Algorithm 2 runs in polynomial time when using our relaxation based algorithm to solve MIN_COST in the main resolution loop.

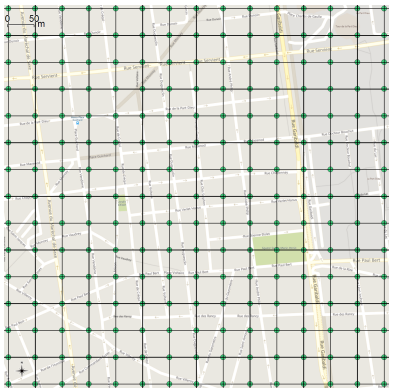
4.4 Performance evaluation

In this section, we present the simulations that we have performed in order to evaluate our proposal. We first present the data set that we used and the common simulation parameters. Then, we provide a proof-of-concept to show how we execute our models on a real dataset. Next, we investigate the performance of the proposed heuristics in terms of solution quality and execution time. After that, we evaluate the coverage results and study the compromise between the pollution mapping quality and the deployment cost. Finally, we assess the impact of pollution variations on the network connectivity while considering the heterogeneous nature of nodes.

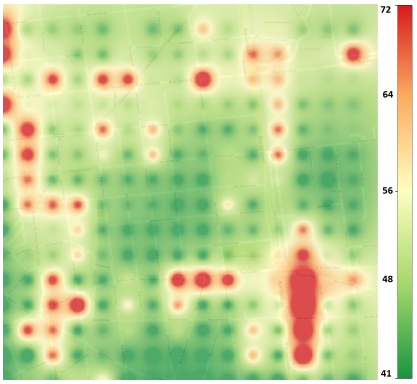
4.4.1 Dataset

In order to consider the real dispersion of air pollutants in the simulated pollution concentrations $\mathcal{M}_p^t$, we perform our simulations on a set of 12 monthly and one annual pollution snapshots. The dataset corresponds to the 2008 Nitrogen Dioxide (NO_2) concentrations in the Lyon district of La-Part-Dieu, which is the heart of the Lyon City. Pollution snapshots are generated by an enhanced atmospheric dispersion simulator called SIRANE [117], which is designed for urban areas and takes into account the impact of street canyons on pollution dispersion. The dataset has been provided by LMFA, which is a research lab specialized in fluid mechanics in the Lyon city.

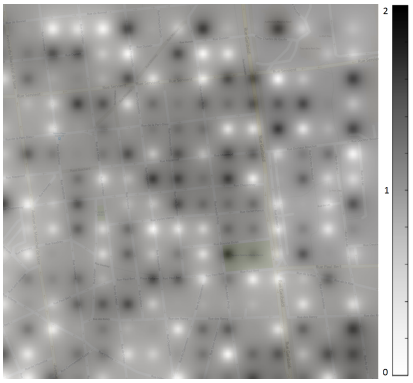
We depict in Fig. 4.1b the pollution map that corresponds to the annual mean of 2008. The deployment region has a spatial resolution of 50 meters and is depicted in



(a) Deployment region



(b) NO_2 concentrations ($\mu g/m^3$)



(c) Simulation errors ($\mu g/m^3$)

Figure 4.1: Deployment region, simulation of 2008 annual concentrations of NO_2 and simulation errors corresponding to the district of La-part-dieu, Lyon.

Fig. 4.1a. We consider as potential positions of nodes all the grid points. In order to show how we take into account simulation errors, we consider a map of randomly generated errors, which is depicted in Fig. 4.1c.

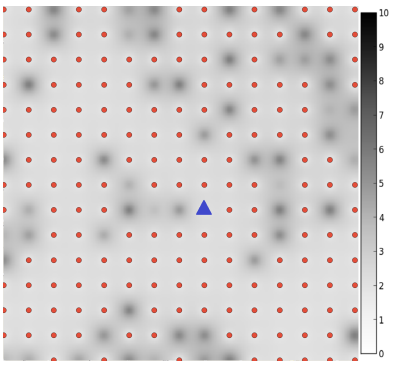
We used as MILP exact solver the IBM ILOG CPLEX solver. Default simulation parameters are summarized in TABLE 4.2. When it is not precised in simulation scenarios, we suppose that sensing is perfect. In addition, we fix the maximum number of sinks to 1 in order to get mono-sink networks.

Parameter	Notation	Value
Number of discrete points	$\mathcal{N}$	225
Maximum correlation distance	d	$100m$
Attenuation coefficient of correlation	α	1
Communication range of sensor nodes	$\mathcal{R}$	$150m$
The tolerated estimation error	$\mathcal{E}$	$5\mu g/m^3$
The maximum number of sinks	$\mathcal{I}$	1
The cost of deploying a sensor at point p	δ_p	1
The cost of deploying a sink at point p	ψ_p	10
Sensing error at point p	s_p	0

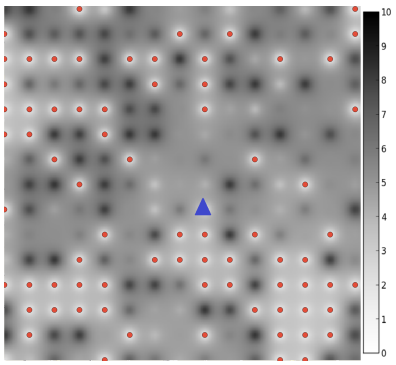
Table 4.2: Default simulation parameters of the air quality mapping approach.

4.4.2 Proof-of-concept: application to the La-Part-Dieu district

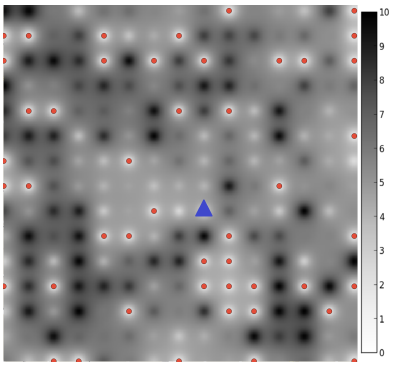
In order to validate our formulation of pollution-aware coverage quality, we run the model MIN_ERROR to minimize the maximum estimation error while considering 3 values for the tolerated estimation error: 5, 8 and 10 $\mu g/m^3$. We depict in Fig. 4.2 the obtained positions of sensors and sinks for the three simulation cases. We also evaluate at each point of the map the estimated concentration and then we calculate the resulting estimation error. The obtained errors are also depicted in Fig. 4.2. We notice that less sensors are used when the tolerated estimation error increases. This is expected since better deployment precision needs more sensor nodes. In addition, Fig. 4.2 shows that the maximum value in the error map in each simulation case is bounded by the tolerated estimation error, which fits our coverage formulation. Moreover, the obtained nodes form a connected network as formulated in our connectivity constraint.



(a) Tolerated error = $5\mu g/m^3$ (181 nodes)



(b) Tolerated error = $8\mu g/m^3$ (85 nodes)



(c) Tolerated error = $10\mu g/m^3$ (58 nodes)

Figure 4.2: Proof-of-concept: Optimal WSN topology and the corresponding estimation errors ($\mu g/m^3$) while considering different values of the maximum tolerated error. Sensors (respectively sinks) are depicted in red circles (blue triangles).

4.4.3 Evaluation of the proposed heuristics

We evaluate the experimental complexity of our optimization models and the performance of our heuristic algorithms. We assume in this simulation case that simulation models of pollution concentrations are perfect, and we focus on solving the model MIN_COST while considering the minimization of the deployment cost of a connected network without exceeding a tolerated estimation error equal to $5\mu g/m3$. TABLE 4.3

presents the results that we obtained while considering the Lyon district of La-part-dieu with different monthly snapshots of pollution concentrations. Results in TABLE 4.3 are sorted according to the execution time of the MILP solver. Note that this sort fits well with maximum pollution variability (i.e. the range length of pollution concentrations corresponding to each snapshot).

TABLE 4.3 shows that the less the pollution variability in the deployment region, the higher is the execution time of the exact MILP solver. For example, the MILP solver takes more than one hour to solve the problem instance corresponding to January. However, the execution time of our heuristic algorithm is stable regarding the pollution variability and is less than one minute in most of the simulations. As a result, the time gain factor of our heuristic increases with the complexity of the problem and is higher than 10 for most of the instances.

As for the value of the objective function, we notice that the optimal deployment cost increases with pollution variability; i.e. more sensors are needed when pollution variability is higher. In order to compare the optimal solutions to the approximated ones given by our heuristic, we define the percentage drift which is equal to $\frac{o_2 - o_1}{o_1}$ where o_1 the optimal objective value and o_2 is the approximated value. We notice in TABLE 4.3 that the percentage drift ranges between 23% and 34% with a mean equal to 30%.

In addition to the comparison to the proposed heuristics, we also analyzed the impact of solving our optimization models in two consecutive steps: we first solve the deployment problem while considering only the coverage constraints and then we run a second time the Cplex solver while considering the connectivity constraints. By solving the problem in this way, the first step provides only a coverage solution while the second step adds the necessary relay nodes in order to ensure the network connectivity. Overall, we noticed in our simulation results that by using this multilevel solving method, the execution time is significantly reduced with a factor exceeding 10 in most of the cases while providing near optimal solutions [35].

4.4.4 Evaluation of the coverage results

In this section, we evaluate the optimal coverage results of our deployment models. First, we depict in Fig. 4.3a the deployment cost depending on the tolerated estimation error while considering two error functions: the maximum error and the mean error. Results show that the less the tolerated error, the higher is the deployment cost, which is reasonable since reducing the estimation error requires more information and hence more sensors. Also, we notice that when considering the same tolerated value on the x-axis, we deploy less sensors for the mean error than the maximum error. Indeed, the maximum error constraint is stronger than the constraint of the mean error.

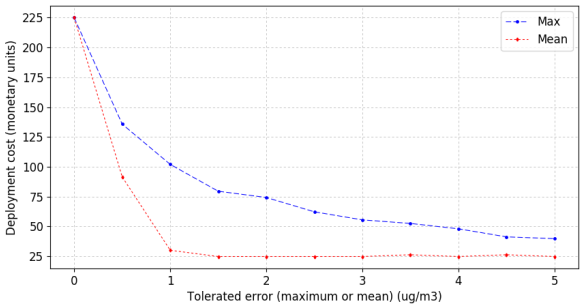
We also evaluate the impact of the sensing (s_p) and simulation (m_p) errors on the final estimation error while considering 4 scenarios, in each of which we vary the values of sensing and simulation errors. We depict the obtained results in Fig. 4.3b. Results show that the estimation error is shifted by $2 * m + s$, which is explained by our definition of the estimation error in formula 4.15 where m occurs two times and s occurs only one time.

Finally, we evaluate the impact of sensing heterogeneity while considering two dif-

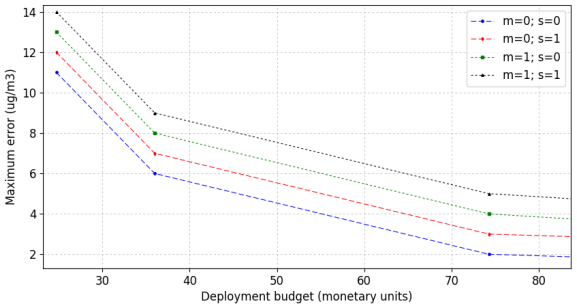
Time (month)	Pollution variability ($\mu g/m^3$)	Execution time (seconds)			Objective function (deployment cost)		
		MILP solver	Heuristic	Gain factor	MILP solver	Heuristic	Drift (%)
September	34.877	183.914	40.850	4.502	95.000	117.000	23.2
April	34.630	188.911	42.459	4.449	92.000	121.000	31.5
March	31.827	471.260	42.534	11.080	84.000	113.000	34.5
November	29.483	822.781	47.396	17.360	76.000	102.000	34.2
February	28.024	1495.090	60.790	24.594	72.000	95.000	31.9
January	27.717	4596.050	47.295	97.178	75.000	100.000	33.3
Mean	32.162	1112.852	46.351	22.843	86.143	111.429	30.0

Table 4.3: MILP solver vs. Linear relaxation-based heuristic.

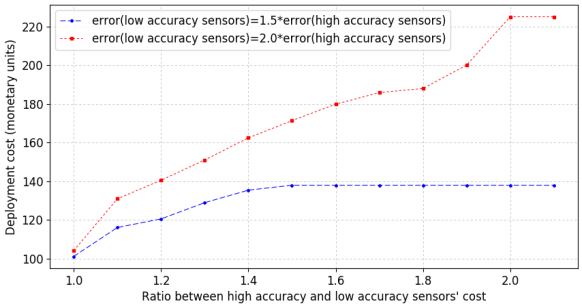
ferent types of sensors: low accuracy and high accuracy ones. We vary the cost ratio of the two sensor types and consider two cases in the quality difference between the two types. Results are depicted in Fig. 4.3c and show that the more the accurate sensors are expensive, the higher is the deployment cost. Indeed, accurate sensors are usually used more than low accuracy sensors.



(a) Deployment cost vs. tolerated error.



(b) Impact of sensing and simulation errors.



(c) Impact of sensing heterogeneity.

Figure 4.3: Impact of the tolerated estimation error and sensing heterogeneity on the deployment cost.

4.4.5 Evaluation of the connectivity results

In the final simulation scenario, we evaluate the connectivity results while considering two different types of nodes depending on their communication capabilities: nodes with a communication range equal to 150m that we consider as short range communication nodes (like 802.15.4 for instance); and nodes with a communication range equal to 500m that we consider as long-range communication nodes (like LoRa or Sigfox for instance). We vary the cost ratio between the two node types and depict the optimal network deployment cost in Fig. 4.4. Results show that only short-range communications are used when long range communications are very expensive; i.e. starting from a ratio equal to the double. Indeed, using several short range communication nodes can replace long range nodes when these latter are much expensive than short range nodes.

4.5 Conclusion

In this chapter, we tackled the deployment issue of heterogeneous sensor networks and proposed mixed integer programming models which take into account the net-

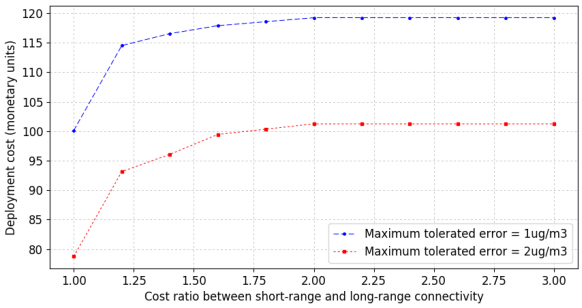


Figure 4.4: Impact of communication heterogeneity on the deployment results.

work deployment cost and the air pollution mapping quality while ensuring the network connectivity. Our main contribution is the definition of an appropriate coverage formulation for air pollution regular mapping in addition to the design of LP-based approximate resolution algorithms. We applied our models and algorithms on a dataset of the Lyon City and evaluated the computational complexity of our proposal. We showed that the less the pollution variability in the deployment region, the higher is the time of solving the optimization models and therefore the higher is the need of using near-optimal heuristics. Moreover, we assessed the impact of the deployment requirements on the coverage and connectivity results. We have shown that the efficiency of a heterogeneous sensor network topology depends on the cost of the different types of sensor nodes.

Chapter 5

Deployment of WSN for Air Pollution Detection

Aiming to construct real-time air pollution maps, the deployment approach presented in the previous chapter ensures that pollution concentrations are either measured or estimated at every point of the deployment region. This is very important when monitoring road-traffic-related pollutants which are present in important concentrations in most parts of urban deployment regions. However, the previous deployment approach is not well adapted to the case of monitoring industry-related very toxic pollutants. Indeed, those pollutants are released in dangerous concentrations mainly because of leaks which happen at factories and industrial sites. Both the spatial and temporal variations of the industry-related pollutants are different compared to road-traffic-pollutants.

In this chapter, we focus on the detection of the threshold crossings of toxic industry-related pollutants and propose an adequate deployment approach which takes into account the locations of potential pollution sources and the physical dynamics of pollution dispersion. The expected deployment should ensure pollution detection and network connectivity while minimizing the deployment cost. We first present an optimization model where pollution coverage and network connectivity are modeled independently. Then, we present an enhanced and more effective model where both coverage and connectivity are jointly formulated using only the flow concept. The enhanced model is compact and tighter than the first one and this allows to reduce the computation time and hence ensures that large-scale problem instances can be solved in reasonable time. The two proposed models take into account the probabilistic sensing of pollution sensors and are designed to handle multiple scenarios of weather conditions. We use a dataset of Greater London in order to evaluate our deployment models while investigating their performance in terms of computational burden. We also study the impact of the nodes' cost, pollution sources density and other deployment parameters on the resulting network topology.

5.1 Background: atmospheric dispersion modeling

We provide in this section an example of the physical pollution propagation models while focusing on the Gaussian model that we use as proof-of-concept in the simula-

tion part of this chapter. The theoretical study of pollution atmospheric dispersion is mainly based on fluid mechanics theory [122]. For the sake of clarity, we focus in this work on steady state dispersion. However, our approach can be used with any other dispersion model that is able to take into account pollution sources like crossroads and highways or the effect of buildings on the dispersion of pollutants.

The basic Gaussian model estimates the concentrations of a pollutant gas released by a pointwise pollution source in a free space environment [123]. The estimated value C (g/m^3) at a measurement location (x, y, z) is given by Formula (5.1). Table 5.1 details the parameters of the model. The pollution source is located at the point $(0, 0, h_s)$ and the measurement point location is given according to a 3D coordinate system where the x-axis is oriented in the wind direction D_w. Parameters σ_y and σ_z describe the stability of the atmosphere and can be approximated using Briggs formulas: $\sigma_y = a_y.|x|^{b_y}$ and $\sigma_z = a_z.|x|^{b_z}$. The parameter H, which represents the pollutant effective release height, is equal to the sum of the pollutant source height h_s and the plume rise Δh. The pollution plume is located above the pollution source and Δh is the vertical distance between the source and the center of the pollution plume. Briggs formulas are commonly used for the calculation of the Δh parameter. To simplify the analysis, we only consider the case where the temperature of the pollutant T_s is greater than the ambient air temperature T, which is usually the case. The value of Δh is given by Formula (5.2) where F, the pollutant gas buoyancy, is computed using Formula (5.3).

$$C(x,y,z) = \frac{Q}{2\pi V_w \sigma_y \sigma_z} e^{-\frac{y^2}{2\sigma_y}} (e^{-\frac{(z-H)^2}{2\sigma_y}} + e^{-\frac{(z+H)^2}{2\sigma_y}}) \tag{5.1}$$

$$\Delta h = \frac{1,6 \cdot F^{1/3} \cdot x^{2/3}}{V_w} \tag{5.2}$$

$$F = \frac{g}{\pi} \cdot V \cdot (\frac{T_s - T}{T_s}) \tag{5.3}$$

The Gaussian model considers only one scenario of weather conditions at a time to compute pollution concentrations. Simulations of the Gaussian model on three scenarios of weather conditions are depicted in Fig. 5.1. Pollution concentrations are given in $\mu g/m^3$ and grouped in 5 intervals. Common values used for simulation are depicted in table 5.2. By comparing Fig. 5.1(a) to Fig. 5.1(b), we notice that wind direction affects the direction of the pollution plume since pollutants are transported by wind. Moreover, variations of the ambient temperature and the wind velocity between Fig. 5.1(b) and Fig. 5.1(c) have affected the concentrations of the pollution plume.

5.2 Deployment models

In this section, we present our deployment models based on integer programming modeling. We base in our coverage formulation on a pollution dispersion model. For the sake of clarity, we use the Gaussian dispersion model in the definition of the pollution inputs of our models. However, these inputs can be also provided by other dispersion models, which may take into account the impact of buildings and urban

Measurement location	
x	Downwind distance from the pollution source (m)
y	Crosswind distance from the pollution source (m)
z	Hight (m)
Pollution emission parameters	
h_s	Pollutant source height (m)
Q	Mass flow rate at the emission point (g/s)
V	Volumetric flow rate at the emission point (m^3/s)
T_s	Pollutant temperature at the emission point (K)
Weather	
T	Ambient air temperature (K)
V_w	Wind velocity (m/s)
D_w	Wind direction (degree)
Constants	
a_y, b_y	Horizontal dispersion coefficients
a_z, b_z	Vertical dispersion coefficients
g	Gravity constant ($9.8m/s^2$)

Table 5.1: Main inputs of the Gaussian dispersion model

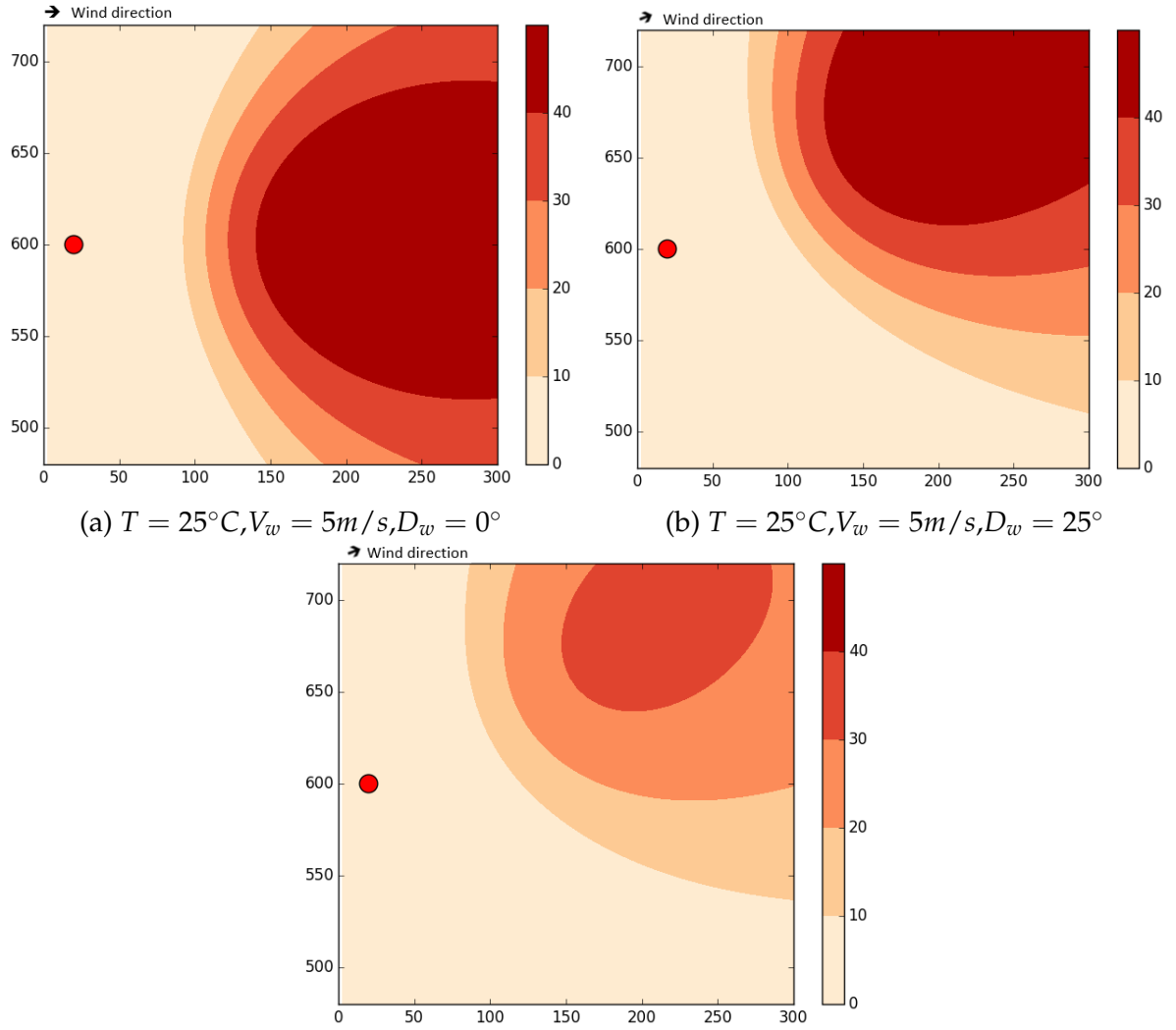
Parameter	value
h_s	$25m$
Q	$5g/s$
V	$1.9mm^3/s$
T_s	30^oC
a_y	1.36
b_y	0.82
a_z	0.275
b_z	0.69

Table 5.2: Simulation parameters of the Gaussian model.

structures. Pollution sources include industrial sources as well as traffic sources such as highways and crossroads. We start this section by presenting the basic model where coverage and connectivity are formulated independently. Then, we present the enhanced model where the two deployment constraints are modeled in a joint way using only the flow concept.

5.2.1 Basic Model

As in the previous chapter, we consider a set of a pre-defined potential positions, denoted $\mathcal{P}$, which is obtained using a discretization of the deployment field restricted to allowed positions. We denote $\mathcal{N} = |\mathcal{P}|$ the number of potential positions. The lo-



(a) $T = 25°C, V_w = 5m/s, D_w = 0°$

(b) $T = 25°C, V_w = 5m/s, D_w = 25°$

(c) $T = 5°C, V_w = 8m/s, D_w = 25°$

Figure 5.1: Simulation of the Gaussian dispersion model with a single pollution source (red point) and different weather conditions.

cations of pollution sources, e.g. factories, sewage treatment plants, crossroads, highways..., is denoted $\mathcal{I}$. $\mathcal{M}$ denotes the number of pollution sources. The binary decision variables x_p, resp. y_p, define if a sensor, resp. a sink, is placed at position p. As in the previous chapter, we denote by δ_p (respectively ψ_p) the deployment cost of a sensor (respectively a sink) at point p. Since a sink embeds sensing capabilities, a sink and a sensor cannot be deployed at the same potential position p as formulated in constraint 5.4. Our optimization models minimize the sensors and sinks overall deployment cost. Thus, we get the objective function in formula 5.5.

$$x_p + y_p \leq 1, \quad p \in \mathcal{P} \tag{5.4}$$

$$\mathcal{F} = \sum_{p \in \mathcal{P}} \delta_p \cdot x_p + \sum_{p \in \mathcal{P}} \psi_p \cdot y_p \tag{5.5}$$

Before tackling the formulation of the coverage and connectivity constraints, we summarize in table 5.3 the notations used in the ILP models.

Sets	
$\mathcal{P}$	Set of potential positions of sensors and sinks
$\mathcal{N}$	Number of sensors and sinks potential positions
$\mathcal{I}$	Set of pollution sources
$\mathcal{M}$	Number of pollution sources
$\mathcal{S}$	Set of weather scenarios
Parameters	
$\mathcal{Z}_i^s$	The pollution zone formed by source i under scenario s
$\mathcal{B}_{ip}^s$	Define whether the position p belongs to the zone Z_i^s or not
$\Gamma(p)$	The neighborhood of the potential position p
δ_p	The cost of deploying a sensor at position p
ψ_p	The cost of deploying a sink at position p
β	Minimum coverage probability to ensure for each zone
$\mathcal{W}_{ip}^s$	The probability of detecting the zone Z_i^s at position p
γ	Percentage of scenarios that have to be taken into account
α_s	Probability that scenario s is realized
C_0	Pollutant concentration threshold
Variables	
x_p	Define whether a sensor is deployed at position p or not $x_p \in \{0,1\},\ p \in \mathcal{P}$
y_p	Define whether a sink is deployed at position p or not $y_p \in \{0,1\},\ p \in \mathcal{P}$
t_i^s	Define whether the zone $\mathcal{Z}_i^s$ is covered or not $t_i^s \in \{0,1\},\ i \in \mathcal{I},\ s \in \mathcal{S}$
g_{pq}	Flow quantity transmitted from node p to node q $g_{pq} \in \{0,1,...\},\ p \in \mathcal{P},\ q \in \Gamma(p)$
f_{ip}^s	Flow quantity transmitted from zone $\mathcal{Z}_i^s$ to node p $f_{ip}^s \in \{0,1\},\ i \in \mathcal{I},\ s \in \mathcal{S},\ p \in \mathcal{Z}_i^s$

Table 5.3: Main notations used in the air pollution detection approach.

Probabilistic pollution coverage

Our aim is to detect the crossings of the threshold C_0 of a given pollutant while covering each pollution source with a probability β. The coverage constraints rely on the modeling of the atmospheric dispersion. We assume that pollution sources release pollutants independently and may have simultaneous release. Our formulation ensures the coverage of threshold crossings in all cases. As shown in section 5.1, pollution concentrations vary depending on weather conditions. Hence, we consider a set of possible weather scenarios $\mathcal{S}$ that can be obtained based on statistical data or weather forecast. A scenario corresponds to a tuple of ambient temperature T, wind velocity V_w and wind direction D_w: $s = (T^s, V_w^s, D_w^s)$. Each scenario s has probability α_s to happen. We assume that $\mathcal{S}$ is a partition of the space of weather conditions, i.e. $\sum_{s\in\mathcal{S}} \alpha_s = 1$ and $s1 \cap s2 = \emptyset,\ \forall s1, s2 \in S$.

Using an atmospheric dispersion model, we determine the set of generated pollution zones. Each zone $\mathcal{Z}_i^s$ corresponds to the geographical area, i.e. set of positions, where the pollution threshold is crossed when the pollution source i is releasing pollutants under the weather scenario s. Let the binary parameter $\mathcal{B}_{ip}^s$ denote whether a position p belongs to $\mathcal{Z}_i^s$ or not. A pollution zone $\mathcal{Z}_i^s$ is therefore the set $\{p \in \mathcal{P}$ *where* $\mathcal{B}_{ip}^s = 1\}$. When using the pointwise Gaussian dispersion model, the value of $\mathcal{B}_{ip}^s$ is calculated using Formula (5.6) where σ_y, σ_z, Q and H are the parameters presented in Section 5.1, $p = (x, y, z)$ and C_0 is the threshold of pollutant concentration above which a point is considered as polluted.

$$\mathcal{B}_{ip}^s = \begin{cases} 1 \text{ if } \dfrac{Q}{2\pi V_w^s \sigma_y \sigma_z} e^{-\frac{y^2}{2\sigma_y}} (e^{-\frac{(z-H)^2}{2\sigma_y}} + e^{-\frac{(z+H)^2}{2\sigma_y}}) \geq C_0 \\ 0 \text{ otherwise} \end{cases} \tag{5.6}$$

A sensor exposed to a given pollutant will detect its concentration with a probability depending on the sensing accuracy. We denote $\mathcal{W}_{ip}^s \in]0, 1[$ the probability of detecting the pollution source i under the weather scenario s at position p, $p \in Z_i^s$. The $\mathcal{W}_{ip}^s$ parameters are mainly due to the technical characteristics of pollution sensors and are not related to the dispersion model.

Once the pollution zones $\mathcal{Z}_i^s$ are identified and the probability parameters $\mathcal{W}_{ip}^s$ are computed, we formulate the coverage of each pollution source i under each weather scenario s with a probability β in constraint 5.7.

$$\prod_{p \in \mathcal{Z}_i^s} (1 - \mathcal{W}_{ip}^s \cdot (x_p + y_p)) \leq (1 - \beta), \quad i \in \mathcal{I},\ s \in \mathcal{S} \tag{5.7}$$

When a sensor or a sink is placed at position p, i.e. $x_p + y_p = 1$, $1 - \mathcal{W}_{ip}^s \cdot (x_p + y_p)$ is then equal to $1 - \mathcal{W}_{ip}^s$, the probability that the node deployed at p do not cover the pollution zone Z_i^s at position p. Assuming that the detection events are independent among all potential positions, constraint 5.7 ensures therefore that each zone Z_i^s is covered with a probability $\beta \in]0, 1[$.

Partial coverage Constraint 5.7 ensures that each pollution source is covered with a probability β under each scenario s. For some applications, we may want to relax this constraint and ask only for the coverage of each pollution source under γ percent of weather scenarios, with a β probability for each scenario. For that, we introduce the binary variable t_i^s that define whether source i is covered during weather scenario s. Therefore $t_i^s = 1$ if a sufficient number of sensors is placed in the pollution zone Z_i^s. The percentage of weather conditions where i can be detected is the sum of the probabilities that a scenario in which i is detected happens. As a result, the coverage formulation of the partial coverage case is given by the constraints 5.8 and 5.9. Constraint 5.8 should be linearized in order to get an ILP formulation. We first apply the log function to get constraint 5.11. Since $x_p + y_p$ and t_i^s are binary, the log can be rewritten to get a linear form as shown in constraint 5.12.

$$\prod_{p \in \mathcal{Z}_i^s} (1 - \mathcal{W}_{ip}^s \cdot (x_p + y_p)) \leq (1 - \beta \cdot t_i^s), \quad i \in \mathcal{I}, s \in \mathcal{S} \tag{5.8}$$

$$\sum_{s \in \mathcal{S}} t_i^s \cdot \alpha_s \geq \gamma, \ \ i \in \mathcal{I} \tag{5.9}$$

$$\log \prod_{p \in \mathcal{Z}_i^s} (1 - \mathcal{W}_{ip}^s \cdot (x_p + y_p)) \leq \log(1 - \beta \cdot t_i^s), \ i \in \mathcal{I}, s \in \mathcal{S} \tag{5.10}$$

$$\sum_{p \in \mathcal{Z}_i^s} \log(1 - \mathcal{W}_{ip}^s \cdot (x_p + y_p)) \leq \log(1 - \beta \cdot t_i^s), \ i \in \mathcal{I}, s \in \mathcal{S} \tag{5.11}$$

$$\sum_{p \in \mathcal{Z}_i^s} (x_p + y_p) \cdot \log(1 - \mathcal{W}_{ip}^s) \leq t_i^s \cdot \log(1 - \beta), \ i \in \mathcal{I}, s \in \mathcal{S} \tag{5.12}$$

Uniform detection probability If the probabilistic sensing values $\mathcal{W}_{ip}^s$ are identical among all the points that are within the same pollution zone $\mathcal{Z}_i^s$, i.e. $\mathcal{W}_{ip}^s = \mathcal{W}_i^s \ \forall p \in \mathcal{Z}_i^s$, constraint 5.7 can be simplified as in 5.13. Using the same binarity argument as before to get the variables out of the log, we get the linear formulation in constraint 5.17.

$$\prod_{p \in \mathcal{Z}_i^s} (1 - \mathcal{W}_i^s \cdot (x_p + y_p)) \leq (1 - \beta), \quad i \in \mathcal{I}, s \in \mathcal{S} \tag{5.13}$$

$$\prod_{p \in \mathcal{Z}_i^s \text{ where } (x_p + y_p = 1)} (1 - \mathcal{W}_i^s) \leq (1 - \beta), \quad i \in \mathcal{I}, s \in \mathcal{S} \tag{5.14}$$

$$(1 - \mathcal{W}_i^s)^{(\sum_{p \in \mathcal{Z}_i^s} (x_p + y_p))} \leq (1 - \beta), \quad i \in \mathcal{I}, s \in \mathcal{S} \tag{5.15}$$

$$(\sum_{p \in \mathcal{Z}_i^s} (x_p + y_p)) \cdot \log(1 - \mathcal{W}_i^s) \leq \log(1 - \beta), \quad i \in \mathcal{I}, s \in \mathcal{S} \tag{5.16}$$

$$\sum_{p \in \mathcal{Z}_i^s} (x_p + y_p) \geq \frac{\log(1 - \beta)}{\log(1 - \mathcal{W}_i^s)}, \quad i \in \mathcal{I}, s \in \mathcal{S} \tag{5.17}$$

Connectivity

As in the previous chapter, we formulate in this deployment model the connectivity constraint as a network flow problem. Let $\Gamma(p), \ p \in \mathcal{P}$ denote the set of sensor p neighbors, and g_{pq} denote the flow quantity transmitted from a node located at potential position p to another node located at potential position q. The following constraints ensure that the deployed sensors and sinks form a connected wireless sensor network; i.e. each sensor can communicate with at least one sink. Constraints 5.18 and 5.19 are designed to ensure that each deployed sensor generates a flow unit in the network

while constraint 5.20 ensures that the deployed sinks cannot transmit flow units, and only act as receivers. The overall flow is conservative thanks to constraint 5.21, which means that the flow sent by the deployed sensors has to be received by the deployed sinks.

$$\sum_{q\in\Gamma(p)} g_{pq} - \sum_{q\in\Gamma(p)} g_{qp} \geq x_p - \mathcal{N}\cdot y_p, \quad p\in\mathcal{P} \tag{5.18}$$

$$\sum_{q\in\Gamma(p)} g_{pq} - \sum_{q\in\Gamma(p)} g_{qp} \leq x_p, \quad p\in\mathcal{P} \tag{5.19}$$

$$\sum_{q\in\Gamma(p)} g_{pq} \leq \mathcal{N}\cdot x_p, \quad p\in\mathcal{P} \tag{5.20}$$

$$\sum_{p\in\mathcal{P}}\sum_{q\in\Gamma(p)} g_{pq} = \sum_{p\in\mathcal{P}}\sum_{q\in\Gamma(p)} g_{qp} \tag{5.21}$$

ILP model

At the end, our general basic optimization model can be written as follows:

[Basic model]
Objective: minimize (5.5)
Pollution coverage constraints: (5.4), (5.9), (5.12)
Connectivity constraints: (5.18), (5.19), (5.20), (5.21)

5.2.2 Enhanced Model

The basic formulation cannot deal with large-scale instances. One of the main reasons is that the two sub-problems, namely connectivity and coverage, are formulated as set of constraints of different natures. To cope with this, we propose in this section a more efficient modeling. By considering pollution sources as a part of the network, we obtain a homogeneous coverage/connectivity formulation as a network flow problem. In the second model, each pollution source i should transmit some flow units to potential nodes p which are located within the pollution zone corresponding to each weather scenario s i.e. $p \in \mathcal{Z}_i^s$. In addition, sensors are flow conservative and the sinks receive the flow units generated by pollution sources. Therefore, the definition of the joint coverage/connectivity is to ensure that sinks will be informed each time that a threshold crossing occurs. In this regard, a sensor has to receive at most one unit from a given pollution zone. We hence define the binary decision variable f_{ip}^s as the flow quantity from the pollution source i to the potential node p in the case of weather scenario s. The following constraints ensure coverage and connectivity for pollution monitoring.

$$\sum_{s \in \mathcal{S}} t_i^s \cdot \alpha_s \geq \gamma, \quad i \in \mathcal{I} \tag{5.22}$$

$$\sum_{p \in \mathcal{Z}_i^s} f_{ip}^s \cdot \log(1 - \mathcal{W}_{ip}^s) \leq t_i^s \cdot \log(1 - \beta), i \in \mathcal{I}, s \in \mathcal{S} \tag{5.23}$$

$$\sum_{i \in \mathcal{I}, s \in \mathcal{S} : p \in \mathcal{Z}_i^s} f_{ip}^s + \sum_{q \in \Gamma(p)} g_{qp} - g_{pq} \leq \mathcal{NM}|\mathcal{S}|y_p, \quad p \in \mathcal{P} \tag{5.24}$$

$$\sum_{i \in \mathcal{I}, s \in \mathcal{S} : p \in \mathcal{Z}_i^s} f_{ip}^s + \sum_{q \in \Gamma(p)} g_{qp} - g_{pq} \geq 0, \quad p \in \mathcal{P} \tag{5.25}$$

$$\sum_{q \in \Gamma(p)} g_{pq} \leq \mathcal{NM}|\mathcal{S}|x_p, \quad p \in \mathcal{P} \tag{5.26}$$

$$f_{ip}^s \leq x_p + y_p, \quad p \in \mathcal{P}, i \in \mathcal{I}, s \in \mathcal{S} \tag{5.27}$$

Coverage is formulated in constraints 5.22 and 5.23. Constraint 5.22 is similar to the basic model and ensures coverage of each pollution source under a γ percentage of weather scenarios. Constraint 5.23 is equivalent to constraint 5.12 of the basic model. Constraint 5.27 enforces that all the flow units are received by deployed nodes. Thanks to constraints 5.24 and 5.25, when a sensor is deployed on point p (case $y_p = 0$ and x_p=1), we ensure that the inflow of sensor p is equal to its outflow; i.e. The flow is conservative on deployed sensors. In addition, constraints 5.24 and 5.25 also ensure that the sinks are allowed to gather all the flow units that are generated in the network (case $y_p = 1$ and x_p=0). Constraints 5.26 and 5.27 combined with constraints 5.24 and 5.25 ensure that absent nodes do not participate in the communication. As a result, the deployed sensors have to send the flow units gathered from pollution sources to the sinks in order to get the connectivity constraints verified. The enhanced optimization model can then be written as follows.

[Enhanced model]
Objective: minimize (5.5)
Coverage-connectivity joint constraints: (5.4), (5.22), (5.23), (5.24), (5.25), (5.26), (5.27)

Particular cases As in the basic model, some simplifications can be made in some particular cases. First, binary variables t_i^s can be avoided when covering all the pollution zones is required, in which case all t_i^s variables will be replaced by 1. Then, when considering homogeneous probabilistic sensing among potential positions of sensors, i.e. $\mathcal{W}_{ip}^s = \mathcal{W}_i^s$ for all $p \in \mathcal{P}$, constraint 5.23 can be transformed to constraint 5.28 using the same transformations done with the basic model.

$$\sum_{p \in \mathcal{P}} f_{ip}^s \geq \frac{\log(1 - \beta)}{\log(1 - \mathcal{W}_i^s)}, \quad i \in \mathcal{I}, s \in \mathcal{S} \tag{5.28}$$

Finally, considering the case of mono sink WSN, constraints 5.24 and 5.25 can be grouped and replaced by the following constraint 5.29.

$$\sum_{i\in\mathcal{I},s\in\mathcal{S}:p\in\mathcal{Z}_i^s} f_{ip}^s + \sum_{q\in\Gamma(p)} g_{qp} - g_{pq} = \sum_{i\in\mathcal{I},s\in\mathcal{S}} K_i^s \cdot y_p, \quad p \in \mathcal{P} \tag{5.29}$$

5.3 Performance evaluation

In this section, we present the simulations that we have performed to evaluate our deployment models. We first present the data set of Greater London that we used in our simulations. Next, as a proof of concept we apply our models to the London Borough of Camden. Then, we investigate the performance of the two optimization models in terms of computational burden. Finally, we study the impact of the sink/sensor cost ratio, nodes' height, pollution sources density, the probabilistic sensing of sensors and weather conditions. This study allows us to derive engineering insights for effective deployments of air pollution sensors in an urban environment.

5.3.1 Greater London dataset

We evaluate our deployment models on a data set provided by the Greater London community [124]. London is one of the most polluted cities in Europe [125]. The data set corresponds to the locations of urban pollution sources. In this data set, mostly urban facilities have the potential to affect the air quality such as petrol stations, waste oil burners, cement works, etc. The set of pollution sources is spread over the 32 boroughs of Greater London. Overall, 1090 pollution sources are considered. Pollution sources are depicted in Fig. 5.2. The distribution of pollution sources per borough depends on the surface of the borough and ranges from 6 sources to 161 sources.

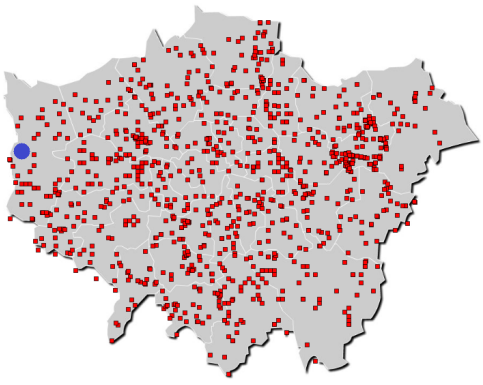


Figure 5.2: Pollution sources (squares) and the weather station (disk) in Greater London.

In addition to pollution sources locations, we compute the weather scenarios leveraging statistical data gathered by a weather station of London [126]. The location of this station is depicted in Fig. 5.2. We consider weather conditions of each month of the

year averaged over the last past 10 years. The set of weather conditions is depicted in table 5.4, each scenario corresponds to values of ambient temperature, wind direction and wind velocity. As proof of concept and without loss of generality, we assume that weather scenarios provided by the considered weather station are homogeneous in all the area of Greater London.

ILP formulations are implemented using the IBM ILOG CPLEX Optimization Studio and executed on a PC with Intel Xeon E5649 processor under Linux. The ILP solver is executed with a time limit of 30 minutes. The default values of simulation parameters are summarized in Table 5.5. We generate the pollution inputs of our deployment models using the Gaussian model presented in section 5.1 while considering the same pollutants characteristics as discussed in section 5.1. We define the nodes neighboring Γ based on a given transmission range. Moreover, we assume that the cost of nodes is independent of the position of the node. Furthermore, we investigate the coverage of pollution sources with respect to all the considered scenarios, i.e. $\gamma = 1.0$. In addition, we consider that the probabilistic sensing value $\mathcal{W}_{ip}^{s}$ is constant and equal to 90%.

5.3.2 Proof-of-concept: application to the London Borough of Camden

As a proof of concept, we first execute our models on the London Borough of Camden. We use streetlights as potential positions of sensors which may alleviate the energy constraints. The streetlights data set was provided by the Camden Datastore [127]. Camden is spread over an area of around $8km \times 6km$ and contains 19 pollution sources. Fig. 5.3 depicts the pollution zones obtained by running the Gaussian dispersion model while taking into account weather conditions of the month of January. Fig. 5.3 also shows the obtained positions of wireless sensor network nodes computed by the deployment models. We notice that sensors are placed at the intersections of the different pollution zones in order to minimize the coverage deployment cost. Moreover, the resulting network consists of 7 sub-networks and a sink is deployed in each one.

The following results have been obtained by running our deployment models on 100 deployment regions that were formed by dividing the Greater London map into a set of $1200m \times 1200m$ blocks. The density of pollution sources varies between 3 and 18 sources per block. We discretize each block with a resolution of $100m$ to get a $2D$ grid of points that we consider as potential positions of WSN nodes.

5.3.3 Tractability evaluation

Compactness and tightness

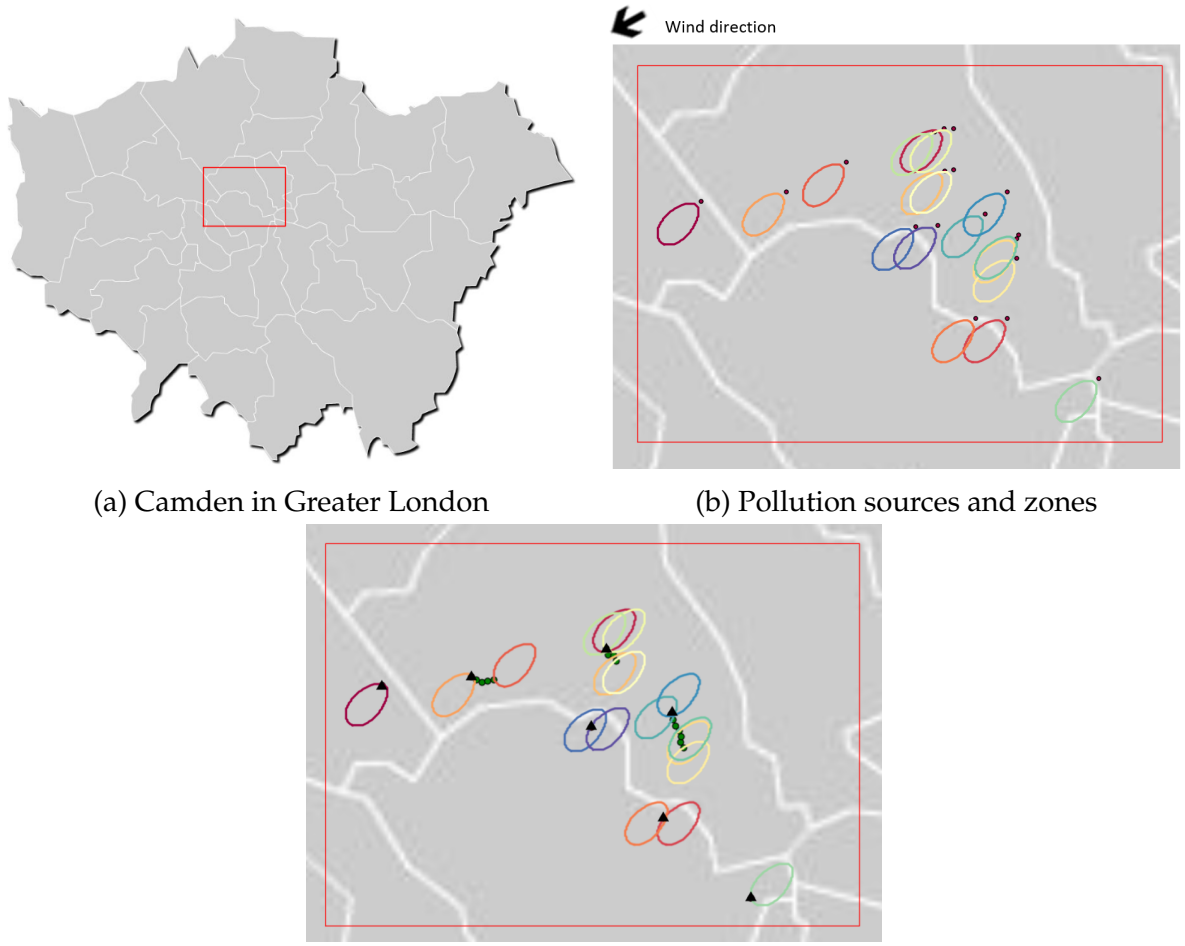
Despite the fact that the models are to run offline, a better model formulation allows the execution on large-scale instances. Integer programming formulations are usually compared in terms of compactness and tightness [128, 129]. The compactness of a model is given by the size of the instances mainly the number of variables and the number of constraints. A compact formulation, i.e. a formulation that allows for small size instances, is not sufficient to get better performances. A good formulation should

Month of year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Ambient temperature (^oC)	7	7	9	13	15	19	21	20	18	14	9	7
Wind velocity (m/s)	5	5	5	5	5	5	5	5	5	5	5	5
Wind direction (degree)	225	247	270	270	270	225	225	225	225	202	225	247

Table 5.4: Weather statistics of London.

Parameter	Value
Nodes transmission range	$100m$
nodes' height	$10m$
Sensors cost (δ_p)	1 monetary unit
Sinks cost (ψ_p)	10 monetary units
Coverage requirements of pollution zones (β)	0.98
Coverage requirements of weather conditions (γ)	1.0
Detection sensitivity of sensors ($\mathcal{W}^s_{ip}$)	0.9
Ambient Temperature (T)	7^oC
Wind velocity (V_w)	$5m/s$
Wind direction (D_w)	225^o
Pollution threshold (C_0)	$20\mu g/m^3$

Table 5.5: Default simulation parameters of the air pollution detection approach.



(a) Camden in Greater London

(b) Pollution sources and zones

(c) Sensors (discs) and sinks (triangles)

Figure 5.3: Application to the London Borough of Camden

be compact and tight at the same time. The tightness of a model allows for a reduced search space and is usually measured using the integrality gap [128]. This is the gap between the optimal objective value of the ILP Z_{ILP} and the optimal objective value of

the relaxed formulation Z_{LP} that is obtained by relaxing the integrality constraints. In the case of a minimization problem, the integrality gap is equal to $(Z_{ILP} - Z_{LP})/Z_{ILP}$. When considering two models with more or less the same compactness, the fastest model is the one that has the lowest integrality gap.

Metric of computational complexity

In order to show the impact of the complexity of the block instances on the tractability of our models, we consider the area of interest as a complexity metric. For a given block b, let C_b be the set of potential positions of sensors that are at least within a pollution zone generated by the block pollution sources under the weather scenarios that are considered. The metric value is defined as the area of the convex envelope of the set C_b. This means that the area of interest includes all the potential positions needed for pollution coverage, i.e. C_b, and also the area where relay nodes may be placed. Indeed, neither coverage sensors nor relay nodes will be placed in the block area that is not included in the area of interest.

Comparison results

As mentioned before, we execute the two deployment models using formulations with respect to the simulation parameters. The two models gave the same objective values; this was expected since the enhanced model is derived from the basic one. We depict in table 5.6 the compactness, the tightness and the execution time of the models depending on the area of interest of block instances. Results have been averaged with respect to the complexity metric class of each instance. We also plot in Fig. 5.4 the distribution of execution time within the classes of the area of interest using a boxplot.

We notice that the instances that are more complex take more time to be resolved when using both of the two models. Moreover, Basic Model and Enhanced Model have nearly the same compactness, the difference is due to f^s_{ip} variables added to the enhanced formulation to define a link between pollution sources and WSN nodes. However, the enhanced model is much tighter than the basic one with lower integrality gaps. This difference in tightness impacts well CPU time and allows to enhance the total mean execution time with a factor of around 8.

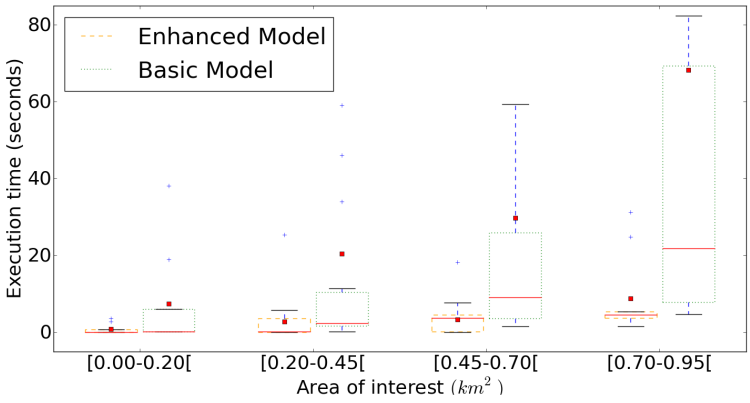


Figure 5.4: Impact of the area of interest on the execution time. Red squares represent the means of the classes and "plus" signs represent the outliers.

Area of interest (km^2)	Compactness (vars. and consts)		Tightness (int. gap)		CPU time (seconds)	
	Basic M.	Enhanced M.	Basic M.	Enhanced M.	Basic M.	Enhanced M.
$[0.00 - 0.20[$	2193	2275	0.880	0.070	7.460	0.890
$[0.20 - 0.45[$	2194	2354	0.870	0.080	20.400	2.810
$[0.45 - 0.70[$	2195	2429	0.850	0.120	29.830	3.360
$[0.70 - 0.95[$	2196	2474	0.830	0.180	68.200	8.820
Mean	**2194.5**	**2383**	**0.860**	**0.110**	**31.470**	**3.970**

Table 5.6: Basic Model vs. Enhanced Model.

We now plot in Fig. 5.5 the histogram of the ratio of execution time between the enhanced model and the basic model. We notice that for around 40% of the block instances, the enhanced model is at least 10 times faster than the basic model. Fig. 5.5 also shows the density function of the ratio of integrality gap between the enhanced model and the basic model. The density function of integrality gap ratio fits well with the histogram of execution time ratio; this is explained by the fact that the integrality gap and the execution time are highly correlated.

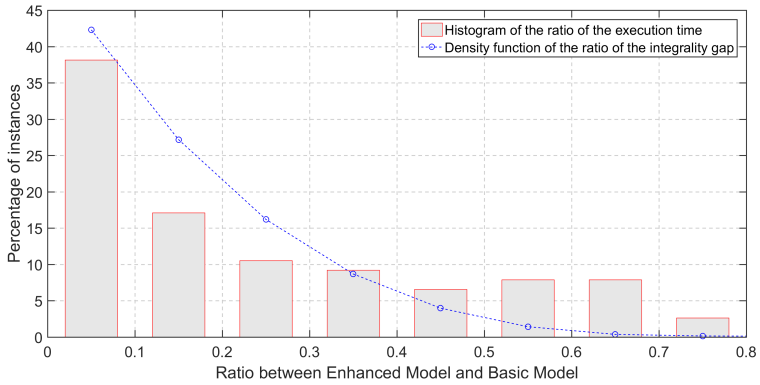


Figure 5.5: Relationship between the integrality gap and the execution time.

Based on the results obtained in this section, we execute only the enhanced model in the following simulations where we investigate the impact of some decision parameters on the deployment results.

5.3.4 Analysis of the network connectivity

In the following, we focus on assessing the network connectivity while evaluating the impact of the nodes' cost and studying the efficiency of communication routes.

Evaluation of the number of nodes

In this simulation case, we analyze the number of sinks and sensors in the resulting networks while varying the ratio between sink cost and sensor cost. We plot in Fig. 5.6 the impact of the cost ratio on the optimal number of sensors and sinks. The cost ratio ranges from 1 to 12 and the results are averaged over all the London blocks defined in the previous section. On one hand, we notice that sensors are less used when their cost is close to the sinks cost. For instance, only sinks are used when the cost ratio is equal to 1. On the other hand, when the cost ratio increases, more sensors are used and the number of required sinks tends to one. As a result, the network is usually formed by only one sink when the cost ratio is greater than 10. This is explained by the fact that adding some relay sensor nodes to ensure connectivity has a less cost than using a lot of sinks that are equipped with pollution sensors.

In the following simulations, we execute the deployment models with a default value of sink/sensor cost ratio equal to 10 as shown in table 5.5. Thus, we use formulations corresponding to the mono-sink case.

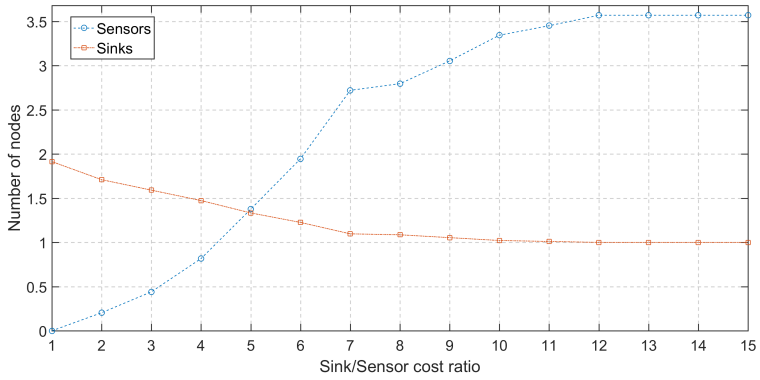


Figure 5.6: Number of nodes depending on the cost ratio.

Evaluation of the number of hops to sink nodes

In this simulation case, we evaluate the obtained networks in terms of the number of hops to sink nodes, which is a measure of the network lifetime and communication delay [130]. As formulated in our connectivity constraints, the positioning of sink nodes does not take into account the length of sensor-to-sink paths. However, sinks can be relocated on the obtained network when the network is monosink, which is the case as shown in the previous subsection. The sink node can be relocated in such a way that the maximum distance to the sink in terms of hops is minimized. This distance is called the radius of the network and describes how much the network is well connected when considering the best position of the sink node. If the sink position is given randomly, the maximum distance to the sink is bounded by the diameter of the network, which is the distance to the sink node when choosing the worst position of the sink.

We depict in Fig. 5.7 the cumulative distribution functions of the network radius and diameter. Results show that the network radius is as at most equal to 5 for more than 96% of the instances. This means that the number of hops to the sink node after relocation is at most equal to 5 in almost all instances.

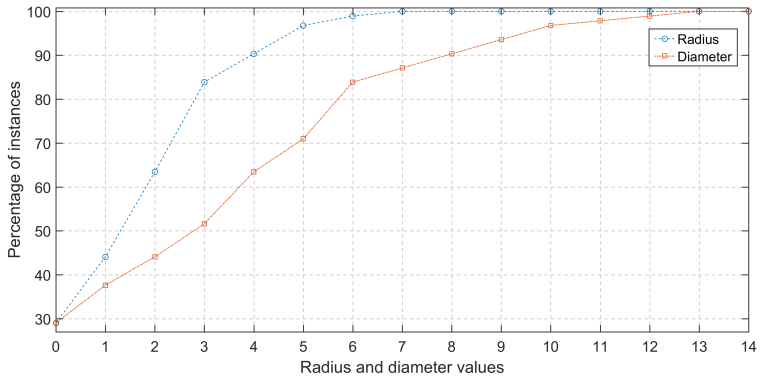


Figure 5.7: Cumulative distribution functions of the network radius and diameter.

5.3.5 Analysis of the coverage results

In the following, we evaluate the deployment results depending on the impact of nodes' height, pollution sources density, the probabilistic sensing of sensors and weather

conditions.

Impact of nodes' height

We now study the impact of the height at which are placed sensors and sinks on the deployment cost. We assume that the height of pollution sources is equal to $25m$, and all the sensors and sinks are deployed at the same height, which is considered in the range from 5 to 40 meters. We plot in Fig. 5.8 the sensors and sinks overall deployment cost depending on their height when applying two different weather scenarios, those corresponding to January and December. The results are averaged over all the London blocks. On one hand, we notice that the deployment cost is minimal when the nodes' height is close to the effective release height of pollution sources H, which is nearly equal to 25.1 in our case. This is explained by the fact that pollution concentration gets the highest values when being near to the pollutant effective release height H. On the other hand, pollutants are more likely to drop than to increase, which is due to gravitation. Indeed, the deployment cost at $40m$ is much greater than the deployment cost at $5m$. Fig. 5.8 also shows that when using different weather scenarios, the deployment cost is not the same. Indeed, weather conditions impacts the disposition of pollution zones allowing for more or less intersections. As a result, the obtained WSN topology depends on the weather conditions taken into account.

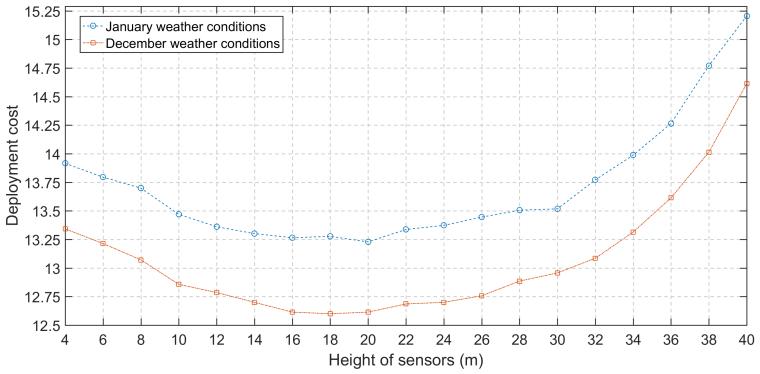


Figure 5.8: Deployment cost average depending on nodes' height with different weather conditions.

Impact of pollution sources density

In this scenario, we study the impact of pollution sources density on the deployment cost. For this purpose, we take the results of the previous scenario corresponding to January weather conditions and averaged with respect to the number of pollution sources of each instance, i.e. the number of pollution sources within each block instance. We plot in Fig. 5.9 the deployment cost variations depending on the nodes' height while considering three different densities: 4, 5 and 6 pollution sources per instance. Fig. 5.9 shows that the more the pollution sources in the environment, the more the number of required sensors and thus the higher is the deployment cost. This can be explained by the number of pollution zones that increases with the number of pollution sources, and thus requires much sensors to ensure the coverage requirements.

In addition, the increasing in the deployment cost from 5 sources density to 6 sources density is less than the increasing from 4 sources density to 5 sources density. This is because when the number of pollution sources increases, more intersections between pollution zones appear and affect the increasing of the deployment cost.

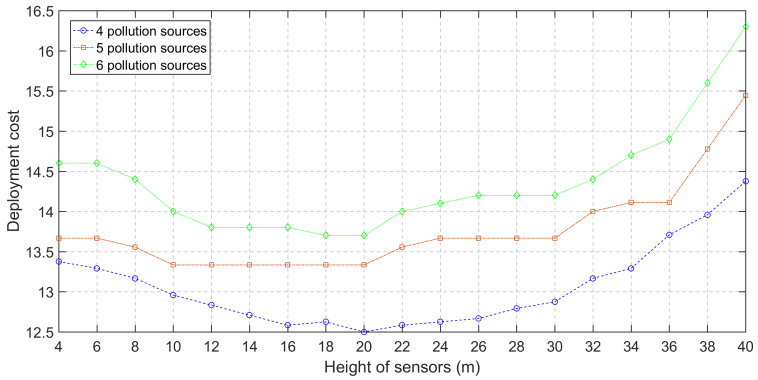


Figure 5.9: Deployment cost average depending on nodes' height and pollution sources density.

Impact of probabilistic sensing

The probabilistic sensing of pollution sensors is one of the most important factors that affect the topology of sensor networks used for pollution monitoring. Fig. 5.10 depicts the average cost of the resulting deployments of the block instances while considering two values of the detection sensitivity of sensors: $\mathcal{W}_{ip}^{s} = 0.9$ and $\mathcal{W}_{ip}^{s} = 0.8$. As expected, using sensors with better detection sensitivity yields less deployment cost. We notice that the ratio between the two curves is around 1.1. This is explained by the intersection existence between the different polluted zones, which means that in some cases a sensor can monitor more than one pollution source.

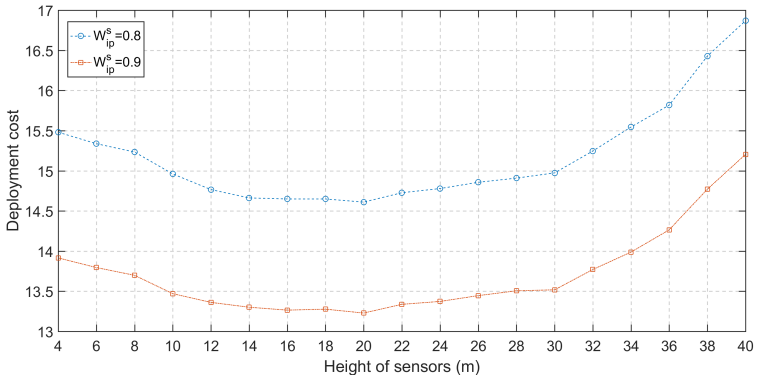


Figure 5.10: Deployment cost average depending on nodes' height and probabilistic sensing values.

Impact of the number of weather conditions

In this simulation case, we study the impact of using a small number of weather scenarios on the deployment results. It is clear that when considering all the possible

weather scenarios, the resulting WSN ensures better pollution monitoring. However, when there is a huge number of weather scenarios, considering a smaller number of these scenarios alleviates the deployment models allowing their application on large-scale instances.

We recall that $\mathcal{S}$ is the set of the monthly weather scenarios presented in table 5.4. Given a subset $\mathcal{S}\prime$ of $\mathcal{S}$, we define the missed pollution zones percentage as the percentage of pollution zones that cannot be covered by the WSN resulting from executing the models under only weather scenarios $S\prime$. Fig. 5.11 illustrates the variations of the missed pollution zones percentage depending on the number of weather scenarios taken into account starting from January weather scenario in the first curve and starting from December weather scenario is the second one. Fig. 5.11 shows that the percentage of missed pollution zones usually decreases when considering more weather scenarios, this is expected since the number of pollution zones depends on the weather scenarios set. However, in some cases, the missed pollution zones percentage remains the same when considering additional scenarios of weather conditions. Indeed, additional weather scenarios involve new pollution zones that may be, in some cases, already included in the set of pollution zones formed without taking into account the additional scenarios. This may happen for instance when additional weather scenarios are slightly different from those already taken into account.

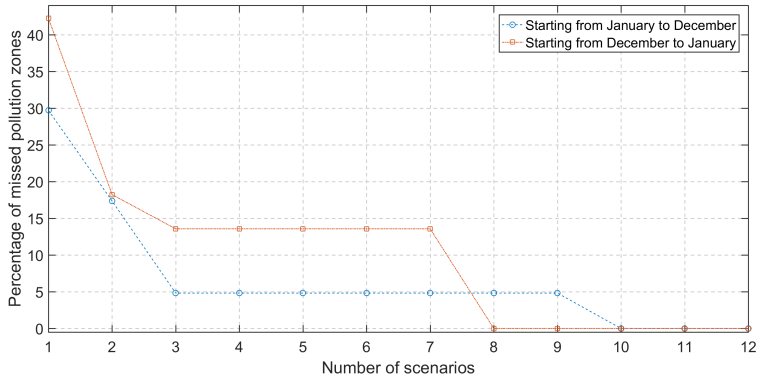


Figure 5.11: Average percentage of missed pollution zones depending on weather scenarios.

In addition to the impact of the number of weather scenarios, their similarity has also to be taken into account. As shown in Fig. 5.11, considering only weather scenarios from December to May meaning only 8 scenarios allows to cover the whole set of pollution zones in contrary to the scenarios set from January to October that requires 10 scenarios.

Impact of the spatiotemporal variation of weather conditions

In this last simulation case, we study the impact of spatial and temporal variations of weather conditions on the deployment results. In order to do so, we use the data of a network of 12 weather stations distributed in all the Greater London [131]. Raw data consist of hourly values of temperature, wind direction (12 possible directions in the range from 0^o to 360^o) and wind velocity of June 2016. We use this data in order

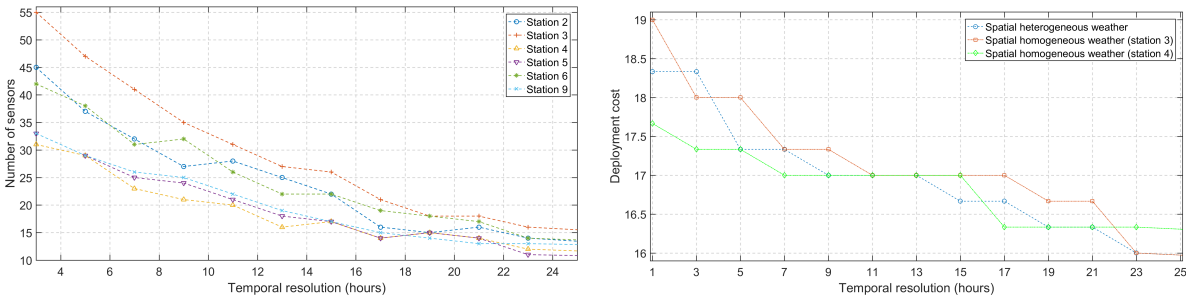
to construct the set of weather scenarios of our deployment model. For each weather station, we have got 672 weather scenarios (28 days and 24 scenarios for each day).

In order to take into account the spatial variation of weather conditions, we propose to consider different weather conditions for each pollution source. When applying the pollution dispersion model on a given pollution source, we use the weather conditions corresponding to the weather station that is the nearest to the given pollution source. Indeed, the Gaussian dispersion model assumes that weather conditions are homogeneous in each pollution zone.

Before the execution of our deployment models, we first eliminate, for each weather station, the weather scenarios that occur more than one time while leaving only one occurrence. We plot in Figure 5.12(a) the number of weather scenarios of some weather stations while varying the aggregation window of data (in the range from 1 hour to 24 hours). We use the mean function for aggregation. We notice that station 3 has the highest temporal variation and station 4 has the lowest one. We did not depict the curves of all the 12 stations for clarity reasons.

We consider in simulation three cases of spatial variation of weather conditions: i) heterogeneous case: we use the data of all the 12 weather stations; ii) 1^{st} homogeneous case: we use only the data of the station that has lowest temporal variation (station 4); iii) 2^{nd} homogeneous case: we use only the data of the station that has highest temporal variation (station 3). We depict in Figure 5.12(b) the mean deployment cost of block instances depending on the temporal aggregation window while considering the 3 spatial cases. Results show that, for all the spatial cases, larger aggregation window involves a smaller number of sensors. This is because the number of weather scenarios is reduced when performing aggregation on larger temporal windows as shown in Figure 5.12(a), i.e. some weather information may be deleted when aggregation is performed on larger temporal windows.

Results also show that on the one hand, when considering the weather data of all the stations (the heterogeneous case), we place more sensors than when considering only the data of the station that has the lowest weather variability (1^{st} homogeneous case, station 4). On the other hand, when considering only the data of the station that has the highest weather variability (2^{nd} homogeneous case, station 3), we place more sensors than when considering the weather data of all the stations (the heterogeneous case). This is due to the fact that station 3 (respectively station 4) involves the maximum (respectively minimum) number of weather scenarios as shown in Figure 5.12(a).



(a) Input (b) Results

Figure 5.12: Impact of the heterogeneity of weather conditions on the deployment results.

5.4 A new pollution-data-aware approach for air pollution detection

In this section, we design a new and different deployment approach which leverages the use of existing pollution datasets. The main difference compared to the optimization solutions previously presented in this chapter is that rather than using the characteristics of pollution sources, we leverage data analysis techniques applied to existing air quality maps in the selection process of the optimal sensor locations. This new deployment approach, which is presented in [34], is based on the spatial analysis of pollution data. In the following, we summarize the main inputs, the workflow and a proof-of-concept of the proposed approach.

5.4.1 Main inputs of the pollution-data-aware approach

In order to deploy sensors efficiently in the city, our approach requires as a primary input, air pollution spatial data. Given a pollutant to monitor, this consists of estimated values of the pollutant concentrations in the whole city for different time instants. Even when the application of pollution dispersion models is not possible, these concentrations can be obtained using interpolation algorithms based on some measurements that are performed by a set of monitoring stations. As in the previously presented deployment approaches, we denote by $\mathcal{T}$ the set of time instants when pollution is estimated, and $\mathcal{P}$ the set of spatial points representing the city. For each time instant $t \in \mathcal{T}$ and spatial point $p \in \mathcal{P}$, let $C_{t,p}$ denote the estimated or measured pollution concentration. In addition to air pollution estimated concentrations, the approach requires also data on sensors' potential positions, which correspond for simplicity sake to a subset of $\mathcal{P}$.

5.4.2 Workflow of the pollution-data-aware approach

The deployment operation is performed through four steps based on the air pollution data and the sensors' potential positions. First, a spatial clustering algorithm is applied to the air pollution data in order to determine pollution zones that are due to the same pollutant sources. To this end, we propose an exploratory algorithm which starts by identifying the points where pollution concentration peaks occur. Then, the neighborhood of each pollution peak is explored to construct the corresponding pollution zone $z \in \mathcal{Z}$.

In the second step, the pollution zones are grouped in order to define a set of deployment sites, denoted by $\mathcal{S}$. Each site $s \in \mathcal{S}$ corresponds to a region in the city and consists of a subset of pollution zones $\mathcal{Z}_s \subset \mathcal{Z}$. We denote the subset of sensors' potential positions that belong to each site $s \in \mathcal{S}$ by $\mathcal{P}_s \subset \mathcal{P}$.

In the last two steps, a mono-sink wireless sensor sub-network is deployed in each site $s \in \mathcal{S}$ so as to cover all the pollution zones $z \in \mathcal{Z}_s$ while minimizing the deployment cost. As a result, the global wireless sensor network is multi-sink. We propose in the third step an integer programming model to deploy a connected sensors' sub-network in each site s while ensuring pollution monitoring. Pollution monitoring is ensured by deploying at least one sensor in each pollution zone in order to ensure the coverage constraints. Finally, in the fourth step, another ILP is proposed in order to

locate a sink node in each deployment site while optimizing the number of hops in sensor-to-sink communication routes.

5.4.3 Proof of concept: application to the Paris city

We evaluate the proposed approach on a pollution dataset of the Paris city while focusing on monitoring NO_2 pollutants. We used pollution data measured by 22 monitoring stations to estimate pollutant concentrations in the whole city. Because air pollution achieved high levels in Paris in March 2014 [132], we decided to use the data of some periods of this month where pollution concentrations were high. Overall, we constructed 10 snapshots of pollution concentrations with a set of 21201 spatial data records per snapshot. In addition to pollution data, we used lamppost locations as sensors' potential positions $\mathcal{P}$.

The first step of the deployment approach is the execution of our spatial clustering algorithm in order to identify the pollution zones where sensors will be deployed. Pollution peaks are first detected and then expanded to form these zones. A number of pollution zones is extracted from each time snapshot. The execution of the spatial clustering algorithm identified 29 pollution zones that occurred in March 2014; these zones are depicted in Fig. 5.13. We notice that some zones occur in different snapshots with different shapes. This is because these zones correspond usually to the same pollutant sources; the different shapes are due to the evolution of weather conditions.

The second step is to group the pollution zones that share intersections into a deployment site where a mono sink sensor network will be deployed. In this simulation case, four sites were identified and are also illustrated in Fig. 5.13. In the third and fourth steps, we execute the sensors' and sinks' optimization models to find the optimal network topology. Fig. 5.13 shows that sensors are placed in the intersections of the pollution zones in order to minimize the financial deployment cost. In addition to sensors placed to ensure pollution zones' coverage, Fig. 5.13 also shows that some sensors are deployed in order to ensure the network connectivity.

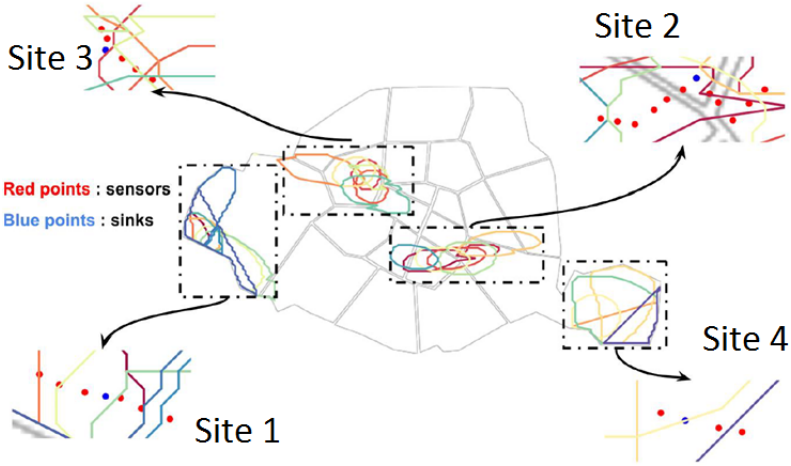


Figure 5.13: Proof-of-concept of the pollution-data-aware approach.

5.5 Conclusion

In this chapter, we focused on the use of wireless sensor networks for the detection of air pollution threshold crossings. We addressed the deployment issue and proposed two optimization models ensuring pollution coverage and network connectivity with the minimum cost. We leveraged the use of atmospheric dispersion modeling to take into account the nature of the addressed phenomenon. Our deployment models are designed in such a way to handle multiple weather scenarios and take into account the probabilistic sensing of nodes. In addition to the coverage formulation based on atmospheric dispersion, we proposed in the second model a joint formulation of coverage and connectivity based only on flows. This allows reducing the computational burden according to our simulation results.

We evaluated the impact of the model parameters on the deployment results. Our results show that sensors should be placed at the effective release height which is close to the one of pollution sources. We also studied the impact of the weather scenarios set on the coverage quality. We showed that this latter depends on the similarity of scenarios, their number, the aggregation temporal window and the spatial heterogeneity of weather conditions. In addition to the weather heterogeneity, the emission variability also affects the deployment results: the higher the emission rates, the less is the number of sensors to deploy.

Chapter 6

WSN Deployment & Scheduling for Air Pollution Simulations' Correction

In the deployment approaches presented in the two previous chapters, we assumed that once the sensors are deployed, air pollution monitoring relies only on the measurements of sensors without taking into account the other possible knowledge sources. Indeed, air pollution maps can be also generated using physical models' simulations. Therefore, sensor measurements can be assimilated in the physical models in order to correct the simulations rather than directly constructing interpolated air pollution maps. In this chapter, we consider this specific application case and tackle first the WSN deployment problem in order to ensure air pollution monitoring through the effective calibration of air pollution physical models. We propose an optimization model based on data assimilation techniques which allow to find the optimal sensor positions that correct in the best way the estimations of the physical models. We also design a heuristic algorithm that is adapted to the deployment problem using linear relaxation. In order to analyze the performance of the deployment approach, we use the dataset of the Lyon city as in chapter 4. Our results show that using data assimilation allows to outperform interpolation-based deployment methods.

In the second part of this chapter, we tackle the scheduling issue of the sensing activity of the already deployed nodes with the objective of ensuring physical models' correction for a maximum period of time by reducing the power consumption of high energy-consuming air pollution sensing probes. We propose a scheduling model to identify the sensors that can be turned off in order to save their energy in case their neighbors measure similar pollution concentrations. We also use the flow concept in order to optimize the communication routes between nodes. We perform a series of simulations while using the energy-consumption characteristics of our lab-designed sensors in order to assess the impact of sensing frequency and transmission power on the network lifetime.

6.1 Problems statement and main inputs

As in the previously presented optimization solutions, we consider as input to the deployment approach the map of a given urban area which is represented by a set of discrete points $\mathcal{P}$. We also use decision variables x_p and y_p to denote the deployment

decisions of sensors and sinks. At a first glance, our objective may seem similar to air quality mapping approaches in the sense that we optimize the deployment of the sensor network by ensuring that for each point $p \in P$, either a sensor is deployed, or pollution concentration can be estimated with a high precision. However, differently from the models presented in chapter 4, we use data assimilation where the estimation of pollution concentrations is based on physical models' simulations in addition to measurements [133]. We estimate pollution concentrations using the drift of physical simulations compared to sensor data instead of interpolating the output measurements of neighboring nodes.

While always focusing on the application of data assimilation for air pollution monitoring, we tackle in a second time the scheduling problem of the already deployed nodes, denoted by the set $\bar{\mathcal{P}}$. Our aim is to schedule the sensing activity of the sensor nodes in order maximize the lifetime of the network while maintaining a good data assimilation quality. Let $\mathcal{T}$ denote the set of time periods over which we would like the network to be operational. We assume that at the end of each time period, sensors send their data to the sink node located at a given point p. Based on the collected data, the sink node estimates the future errors of the physical model at each point of the region of interest. In order to save the energy of sensor nodes, we aim at turning off the nodes located where the physical model future estimated errors do not exceed a critical level.

We use binary decision variables $\bar{x}_p^t$ to specify if sensor p is scheduled to be active during the time period t or not. Our objective is therefore to determine for each future time period $t \in \mathcal{T}$ the state of sensors (ON for active mode and OFF for sleep mode) based on the already collected data and the threshold of the estimation error. This means that when the sensor located at point p is turned off, the data of the neighboring active sensors should be sufficient enough to correct the physical model output at point p. In addition to coverage constraints, the network should remain connected during each time period in order to ensure the communication of sensor data to the sink node. We summarize the main notations of both the deployment and the scheduling approaches in table 6.1.

6.2 Data assimilation formulation

When assimilating sensor measurements in the simulations of physical models, the estimated concentration value $\widehat{\mathcal{Z}}_p^t$ at a given location p where no sensor data are available is formulated as the sum of $\mathcal{M}_p^t$, which is the physical model simulation value at p, and a weighted combination of the difference between the physical model values $\mathcal{M}_q^t$ and the measured values at available neighboring sensor nodes $\mathcal{Z}_q^t$ [134]. For presentation's sake, we use in this section a_q variables to denote the fact that a neighboring sensor is available or not at point q when performing the data assimilation estimation; i.e. a_q is equal to x_q in the deployment problem while when dealing with the scheduling problem $a_q = \bar{x}_q^t$.

As in the interpolation-based estimation, the weights used for the estimation are denoted by $\mathcal{W}_{pq}$. Without loss of generality, we focus on the case of deterministic data assimilation, in which case $\widehat{\mathcal{Z}}_p^t$ is calculated using formula 6.1 [134].

Deployment/scheduling common inputs	
$\mathcal{G}_p^t$	Ground truth pollution concentrations (unknown)
$\mathcal{Z}_p^t$	Measured pollution concentrations (unknown)
$\mathcal{M}_p^t$	Simulated pollution concentrations (using physical models)
$\widehat{\mathcal{Z}}_p^t$	Estimated pollution concentrations (using data assimilation)
m_p^t	Simulation errors
s_p^t	Sensing errors
$\mathcal{W}_{pq}$	Correlation coefficients
$\mathcal{D}$	The correlation distance function
$\Gamma(p)$	Communication neighborhoods
$\mathcal{R}$	Communication range
E	Required assimilation variance
Deployment specific inputs	
$\mathcal{P}$	Set of points approximating the region of interest ($\lvert\mathcal{P}\rvert = N$)
δ_p	The cost of sensors
ψ_p	The cost of sinks
x_p	Define whether a sensor is to be deployed at point p or not ; $x_p \in \{0,1\},\ p \in \mathcal{P}$
y_p	Define whether a sink is to be deployed at point p or not; $y_p \in \{0,1\},\ p \in \mathcal{P}$
Scheduling specific inputs	
$\bar{\mathcal{P}}$	Locations of already-deployed sensors ($\lvert\bar{\mathcal{P}}\rvert = \bar{\mathcal{N}}$)
$\mathcal{T}$	Set of time periods
EI_p	Battery capacity
ES_p^t	Energy consumption due to sensing
ET_p^t	Energy consumption due to transmission
ER_p^t	Energy consumption due to reception
y_p	Define whether a sink is already deployed at point p or not
$\bar{x}_p^t$	Define whether sensor p is active during t or not ; $x_p^t \in \{0,1\}$
θ_t	Define whether the network is operational during time period t or not ; $\theta_t \in \{0,1\}$

Table 6.1: Main notations used in the physical models' correction solutions.

$$\widehat{\mathcal{Z}}_p^t = \mathcal{M}_p^t + \frac{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot a_q \cdot (\mathcal{Z}_q^t - \mathcal{M}_q^t)}{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot a_q} \tag{6.1}$$

We recall that $\mathcal{G}_p^t$ denotes the ground truth (or real) concentration value at point p. In addition, m_p^t (respectively s_p^t) denotes the physical model error (respectively the

sensing error of nodes) which is defined as $\mathcal{M}_p^t - \mathcal{G}_p^t$ (respectively $\mathcal{Z}_p^t - \mathcal{G}_p^t$). With these definitions, formula 6.1 can be transformed into formula 6.2.

$$\widehat{\mathcal{Z}}_p^t = \mathcal{M}_p^t - \frac{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot a_q \cdot (m_q^t - s_q^t)}{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot a_q} \tag{6.2}$$

The data assimilation equation in formula 6.2 is constrained by formula 6.3, which ensures that the denominator is never equal to 0. $\mathcal{B}_{pq}$ parameters define whether there is a correlation between points p and q or not; that is, $\mathcal{B}_{pq} = 1$ when $\mathcal{W}_{pq} > 0$.

$$\sum_{q \in \mathcal{P}} \mathcal{B}_{pq} \cdot a_q \geq 1 \tag{6.3}$$

Given the formula of the assimilation estimated value $\widehat{\mathcal{Z}}_p^t$, the assimilation error with respect to the ground truth value (i.e. $\widehat{\mathcal{Z}}_p^t - \mathcal{G}_p^t$) can be derived as in formula 6.4.

$$\mathcal{E}_p^t = m_p^t - \frac{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot a_q \cdot (m_q^t - s_q^t)}{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot a_q} \tag{6.4}$$

Note that both physical model simulation errors (m_p^t and m_q^t) and sensing errors (s_q^t) are unknown values. Therefore, we propose in this work to consider these errors as random variables where only the variance and the expectation are known by means of empirical analysis of the already collected data. We assume that the expectation of the errors is equal to 0. This is not a strong assumption since both the physical model and sensors can be calibrated to get an error expectation equal to 0 by adding or subtracting the real expectation. That is, the variance defines how much the model (or the sensors) are incorrect at a given point. Based on these assumptions, we define the coverage quality at a given point p as the variance of the assimilation error. To get this formulation, we apply the variance function to formula 6.4 while assuming that sensing errors are independent between themselves and are also independent with respect to the physical model errors. Hence, we get formula 6.5 where Var (respectively Cov) denotes the variance (respectively covariance) function.

$$\begin{aligned} Var(\mathcal{E}_p^t) = Var(m_p^t) &+ \frac{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq}^2 \cdot a_q \cdot (Var(m_q^t) + Var(s_q^t))}{(\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot a_q)^2} \\ &- 2 \cdot \frac{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot a_q \cdot Cov(m_p^t, m_q^t)}{\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot a_q} \\ &+ \frac{\sum_{q_1 \neq p} \sum_{q_2 \neq p, q_1} \mathcal{W}_{pq_1} \cdot \mathcal{W}_{pq_2} \cdot a_{q_1} \cdot a_{q_2} \cdot Cov(m_{q_1}^t, m_{q_2}^t)}{(\sum_{q \in \mathcal{P}} \mathcal{W}_{pq} \cdot a_q)^2} \end{aligned} \tag{6.5}$$

Note that the covariance $Cov(m_p^t, m_q^t)$ is mathematically a function of correlations $\mathcal{W}_{pq}$ and variances $Var(m_p^t)$ and $Var(m_q^t)$ as in formula 6.6 [135].

$$COV(m_p^t, m_q^t) = \mathcal{W}_{pq} \cdot \sqrt{VAR(m_p^t) \cdot VAR(m_q^t)} \tag{6.6}$$

In order to ensure that the air pollution monitoring requirements are met, the coverage quality presented in formula 6.5 should not exceed a threshold (or required) value E as follows:

$$\begin{aligned} & Var(m_p^t) + \frac{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq}^2 \cdot a_q \cdot (Var(m_q^t) + Var(s_q^t))}{(\sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot a_q)^2} \\ & -2 \cdot \frac{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot a_q \cdot Cov(m_p^t, m_q^t)}{\sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot a_q} \\ & + \frac{\sum_{q_1\neq p} \sum_{q_2\neq p, q_1} \mathcal{W}_{pq_1} \cdot \mathcal{W}_{pq_2} \cdot a_{q_1} \cdot a_{q_2} \cdot Cov(m_{q_1}^t, m_{q_2}^t)}{(\sum_{q\in\mathcal{P}} \mathcal{W}_{pq} \cdot a_q)^2} \\ & \leq E \end{aligned} \tag{6.7}$$

We seek in the following sections of this chapter linear optimization models and therefore, we need to linearize constraint 6.7 by eliminating the fraction and the multiplications between the decision variables. We first multiply both sides of formula 6.7 by the denominator of the fraction. Next, we simplify the parts where the square function is applied to variables a_q. Hence, we obtain the linear form of our coverage formulation in formulas 6.8 and 6.9 where expressions $expr_1$ and $expr_2$ are detailed in formulas 6.10 and 6.11 respectively. Finally, real variables $v_{q_1q_2}^t$ correspond to the linear form of the product of decision variables a_{q_1} and a_{q_2} thanks to constraints 6.12.

$$\begin{aligned} & (Var(m_p^t) - E) \cdot expr_1 \\ & + \sum_{q\in\mathcal{P}} \mathcal{W}_{pq}^2 \cdot a_q \cdot (Var(m_q^t) + Var(s_q^t)) \\ & -2 \cdot expr_2 \\ & + \sum_{q_1\neq p} \sum_{q_2\neq p, q_1} \mathcal{W}_{pq_1} \cdot \mathcal{W}_{pq_2} \cdot v_{q_1q_2}^t \cdot Cov(m_{q_1}^t, m_{q_2}^t) \\ & = \mathcal{C}_p^t,\ p \in \mathcal{P} \end{aligned} \tag{6.8}$$

$$\mathcal{C}_p^t \leq 0,\ p \in \mathcal{P} \tag{6.9}$$

$$expr_1 = \sum_{q_1\in\mathcal{P}} \sum_{q_2\in\mathcal{P}} \mathcal{W}_{pq_1} \cdot \mathcal{W}_{pq_2} \cdot v_{q_1q_2}^t \tag{6.10}$$

$$expr_2 = \sum_{q_1\in\mathcal{P}} \sum_{q_2\in\mathcal{P}} \mathcal{W}_{pq_1} \mathcal{W}_{pq_2} v_{q_1q_2}^t Cov(m_p^t, m_{q_1}^t) \tag{6.11}$$

$$\begin{aligned} & v_{q_1q_2}^t \leq a_{q_1}, \quad q_1, q_2 \in \mathcal{P} \\ & v_{q_1q_2}^t \leq a_{q_2}, \quad q_1, q_2 \in \mathcal{P} \\ & v_{q_1q_2}^t \geq a_{q_1} + a_{q_2} - 1, \quad q_1, q_2 \in \mathcal{P} \end{aligned} \tag{6.12}$$

6.3 WSN deployment model

In this section, we use integer programming modeling to derive an optimization model for the deployment of WSN nodes based on the formulation of the assimilation error that we presented in the previous section. The proposed deployment model allows us to minimize the overall deployment cost of sensor and sink nodes in order to guarantee a given target assimilation error while ensuring the connectivity of the network.

6.3.1 Deployment cost

While taking into account the costs of sensors and sinks (i.e. δ_p and ψ_p), the objective function to minimize corresponds to the network overall deployment cost and is defined as follows:

$$\text{Minimize} \sum_{p \in \mathcal{P}} \delta_p \cdot x_p + \sum_{p \in \mathcal{P}} \psi_p \cdot y_p \tag{6.13}$$

6.3.2 Air pollution coverage

As already stated, we propose in our deployment model to ensure the required coverage quality by placing the sensors in such way that the variance of the assimilation error is less than the required variance E. To that end, we use the linear constraints (6.3), (6.8), 6.9, (6.10), (6.11), (6.12) that we designed in the previous section.

6.3.3 Network connectivity

As in the previous contribution chapters, we use the flow concept to ensure the network connectivity. We recall that $\Gamma(p),\ p \in \mathcal{P}$ is the set of neighbors of a node located at point p and g_{pq} are the flow variables. Sensors generate flow units that are later captured by sink nodes thanks to constraints 6.14, 6.15, 6.16, 6.17 and 6.18.

$$\sum_{q \in \Gamma(p)} g_{pq} - \sum_{q \in \Gamma(p)} g_{qp} \geq x_p - (\mathcal{N} + 1) \cdot y_p, p \in \mathcal{P} \tag{6.14}$$

$$\sum_{q \in \Gamma(p)} g_{pq} - \sum_{q \in \Gamma(p)} g_{qp} \leq x_p, \quad p \in \mathcal{P} \tag{6.15}$$

$$\sum_{q \in \Gamma(p)} g_{pq} \leq N \cdot x_p, \quad p \in \mathcal{P} \tag{6.16}$$

$$\sum_{p \in \mathcal{P}} \sum_{q \in \Gamma(p)} g_{pq} = \sum_{p \in \mathcal{P}} \sum_{q \in \Gamma(p)} g_{qp} \tag{6.17}$$

$$\sum_{p \in \mathcal{P}} y_p \leq \mathcal{I} \tag{6.18}$$

6.3.4 Deployment model

We present in what follows the optimization model where we minimize the overall deployment cost subject to coverage and connectivity constraints. It is noteworthy that we can also consider the dual problem where we optimize coverage quality by minimizing the assimilation variance subject to a given deployment budget which should not be exceeded.

Objective: (6.13)
Pollution coverage constraints: (6.3), (6.8), (6.9), (6.10), (6.11), (6.12)
Connectivity constraints: (6.14), (6.15), (6.16), (6.17), (6.18)
Decision variables: $x_p,\ y_p \in \{0,1\}$
Auxiliary variables: $v_{q_1q_2} \in [0,1],\ g_{pq} \in \mathbb{R}^+$

6.4 WSN scheduling model

Based on our data assimilation formulation, we present in this section an efficient WSN scheduling optimization model. We ensure that the WSN network is operational during a maximum number of time periods while guaranteeing the coverage requirements in addition to network connectivity.

6.4.1 Coverage requirements

First, we denote in what follows by binary variables θ_t the fact that the network is operational during time period t or not. Based on the coverage formulation presented in formulas 6.8 and 6.9, we get the following constraint while integrating θ_t variables. Here, M is a sufficiently big number used to ensure that there is always a feasible solution in the case where the network cannot be operational during the whole number of time periods.

$$C_p^t \leq (1-\theta_t)\cdot M \tag{6.19}$$

6.4.2 Connectivity constraints

While using the flow concept, we formulate the connectivity constraints similarly to the deployment model presented in the previous section. In addition to g_{pq}^t flow variables, and due to the fact that the sink node is assumed to be active all the time, we also use variables f_{pq}^t to denote the flow quantity transmitted from a sensor located at point p to a sink node located at point q. The following constraints ensure that the active sensors form together with sink nodes a connected wireless sensor network; i.e. each sensor can communicate its flow units to sink nodes.

$$\sum_{q\in\Gamma(p)} g_{pq}^t + \sum_{q\in\Gamma(p)} f_{pq}^t - \sum_{q\in\Gamma(p)} g_{qp}^t = \bar{x}p^t \tag{6.20}$$

$$\sum_{q\in\Gamma(p)} g_{pq}^t + f_{pq}^t \leq \bar{N}\cdot \bar{x}p^t \tag{6.21}$$

$$\sum_{q\in\Gamma(p)} f_{pq}^t \leq \bar{N}\cdot y_q \tag{6.22}$$

$$\sum_{p\in\bar{\mathcal{P}}}\sum_{q\in\Gamma(p)} f_{pq}^t = \sum_{p\in\bar{\mathcal{P}}} \bar{x}p^t \tag{6.23}$$

6.4.3 Energy consumption constraints

First, let EI_p, ES_p^t, ET_{pq}^t and ER_{pq}^t denote respectively the initial amount of energy (battery capacity) of sensor p, the energy consumption due to sensing during time period t, the energy consumption due to the transmission to a neighbor during time period t and the energy consumption due to the reception from a neighbor during time period t. The following constraints ensure that sensor nodes cannot consume more than their initial amount of energy.

$$\forall p \in \mathcal{P} \ \sum_t ES_p^t \cdot x_p^t + \sum_{t,q} ET_{pq}^t \cdot g_{pq}^t + \sum_{t,q} ER_{pq}^t \cdot g_{qp}^t \leq EI_p \tag{6.24}$$

6.4.4 Lifetime of the network

Finally the network lifetime to maximize corresponds to the number of time periods during which the network is operational as follows:

$$Maximize \sum_{t \in \mathcal{T}} \theta_t \tag{6.25}$$

$$\alpha_0 \geq \alpha_1 \geq \alpha_2 \geq ... \geq \alpha_{|\mathcal{T}|} \tag{6.26}$$

6.4.5 Scheduling model

We present in the following our scheduling model where we maximize the lifetime of the network subject to coverage and connectivity constraints.

Objective: (6.25)
Pollution coverage constraints: (6.3), (6.8), (6.10), (6.11), (6.12), (6.19), (6.26)
Connectivity constraints: (6.20), (6.21), (6.22), (6.23)
Decision variables: $\bar{x}_p \in \{0, 1\}$
Auxiliary variables: $v_{q_1 q_2}^t \in [0, 1]$, $g_{pq}^t, f_{pq}^t \in \mathbb{R}^+$

6.5 Resolution of the models

Since the proposed deployment and scheduling models are based on integer linear programming, solving them using exact solvers takes exponential time. We recall that what the makes MILP models difficult to solve is the number of binary variables which causes an exponential increase in the number of iterations when using the exact MILP solvers. Therefore, the complexity of the MILP models presented in this chapter is mainly due to the number of sensor locations and time periods. As in chapter 4, and in order to solve the optimization models on large instances in a reasonable time while getting good solutions, we use the concept of linear relaxation. For each of our models, we first remove the integrality constraints in order to get relaxed LP models. This

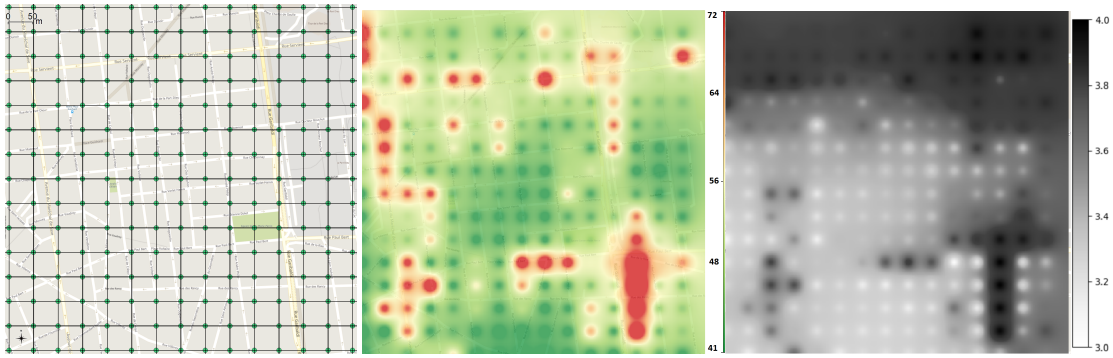
means that binary variables are considered real in the range of $[0, 1]$ and hence the solutions of the LP model are not necessarily binary. We leverage the fact that in a given solution of the LP models where binary variables are fractional, the variable having the maximum value corresponds to the most important variable in the satisfaction of the model constraints. We therefore run iteratively the LP models while fixing to 1 the most important fractional variable at the end of each iteration. This way, we get our deployment or scheduling solution once there are no fractional variables at the end of the LP loop.

6.6 Performance evaluation of the deployment approach

In this section, we present the simulations that we have performed in order to evaluate our data-assimilation-aware deployment approach. We first present the data set that we used and the common simulation parameters. Then, we provide a proof-of-concept while applying the optimization approach to La-Part-Dieu district in Lyon. Next, we compare our proposal to interpolation-based deployment and evaluate the coverage results. Finally, we assess the impact of pollution estimation requirements on the network connectivity.

6.6.1 Dataset

We perform the evaluation of our proposal on monthly pollution data corresponding to the 2008 Nitrogen Dioxide (NO_2) concentrations in the Lyon district of La-Part-Dieu. To illustrate the pollution data set, we depict in Fig. 6.1b a pollution map that corresponds to the annual mean of 2008. The region of interest has a spatial resolution of 50 meters and is depicted in Fig. 6.1a. We consider as potential positions of nodes all the grid points (225 in total). We calculate the correlation coefficients $\mathcal{W}_{pq}$ using an exponential decay function. That is, the correlation between points decreases exponentially with the Euclidean distance.



(a) (b) (c)

Figure 6.1: Region of interest, simulation of 2008 annual concentrations of NO_2 ($\mu g/m^3$) and simulation errors variance ($\mu g^2/m^6$) corresponding to the district of La-part-dieu, Lyon.

We recall that the main input of our deployment approach is the variance of the

errors of the physical model. In this evaluation part, we assume that the errors of the model are linearly correlated with its concentrations. Let γ express the linear relationship between the model concentrations and the model errors. Thus, we first calculate the variance of the concentrations of the physical model based on the 12 monthly pollution maps and then we multiply these variances by γ^2 to get the variance of the physical model errors. We calculate the γ parameter by evaluating the linear regression between the concentrations of the dataset of the physical model and the real data of the few monitoring stations which are already deployed in the Lyon city. The resulting variance map of the physical model errors is depicted in Fig. 6.1c. Default simulation parameters are summarized in table 6.2.

Parameter	Notation	Value
Number of discrete points (region of interest)	$\mathcal{N}$	225
Communication range of sensor nodes	$\mathcal{R}$	$100m$
The maximum number of sinks	$\mathcal{I}$	1
The cost of deploying a sensor at point p	δ_p	1
The cost of deploying a sink at point p	ψ_p	10

Table 6.2: Default deployment parameters of the physical models' correction approach.

6.6.2 Proof-of-concept: application to the La-Part-Dieu district

In order to provide a proof of concept of our assimilation-based coverage formulation, we consider 3 different values for the deployment budget and evaluate the assimilation error provided by the sensor network that is generated by our model. We first consider only the coverage constraints to get the positions of sensor nodes and then we add the connectivity constraints to obtain the positions of the sink and relay nodes which are used only for connectivity. We depict in Fig. 6.2 the positions of sensors, relay nodes and sinks for the three simulation cases. Sensors are placed near streets because these are heavily polluted areas and therefore have the most of uncertainty in physical models.

We also evaluate at each point of the map the corresponding assimilation error. We notice that the assimilation error is reduced when providing higher deployment budget. This is expected since better deployment precision requires more sensor nodes. In addition, Fig. 6.2 shows that the obtained nodes form a connected network as formulated in our connectivity constraint.

6.6.3 Comparison to interpolation-based deployment

In this simulation case, we compare the data-assimilation-aware optimization model to our interpolation-based deployment model which minimizes the estimation error of pollution concentration based only on sensor measurements [35]. We vary the deployment budget and then execute the data-assimilation-aware model based on the variance of the errors and the correlation between the points. We then execute the interpolation-based model by taking the physical model concentrations as a reference.

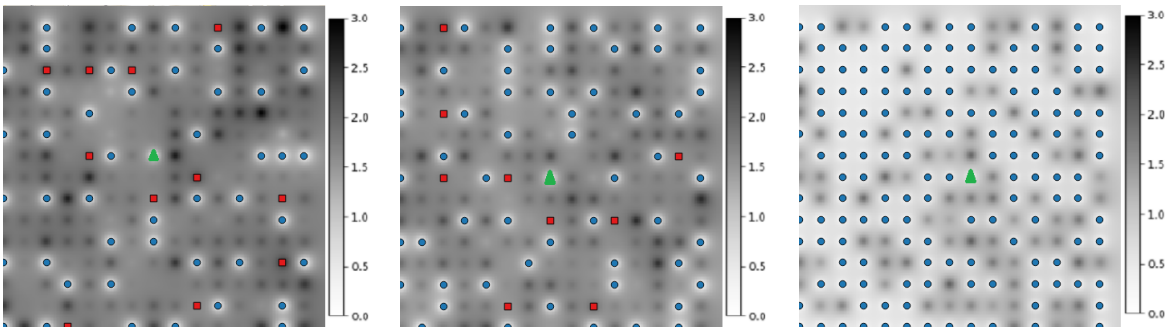


Figure 6.2: Proof-of-concept of the deployment model: optimal WSN topology and the corresponding estimation errors' variance ($\mu g^2/m^6$) while considering different values of the deployment budget (from left to right: respectively 68, 75 and 155 monetary units). Sensors (respectively relay nodes and sinks) are depicted in blue circles (respectively red squares and green triangles). Note that the scale in these 3 figures is different than the scale of Fig. 6.1c.

Once we get the deployment result for each approach, we evaluate the estimation error by running 100 simulations, in each simulation the model errors are considered Gaussian. We then calculate based on these 100 simulations the estimation error variance maps. Finally, we plot the average of the estimation error variance over all the points for each value of the deployment budget in Fig. 6.3.

Results show that the assimilation approach gives better estimation compared to the interpolation approach. This is mainly due to minimizing the variance of the estimation errors in the optimization process rather than trying to get interpolation results that resemble the physical model. Moreover, the difference between the two approaches decreases as the deployment budget increases since the estimation is less used when more sensors are available.

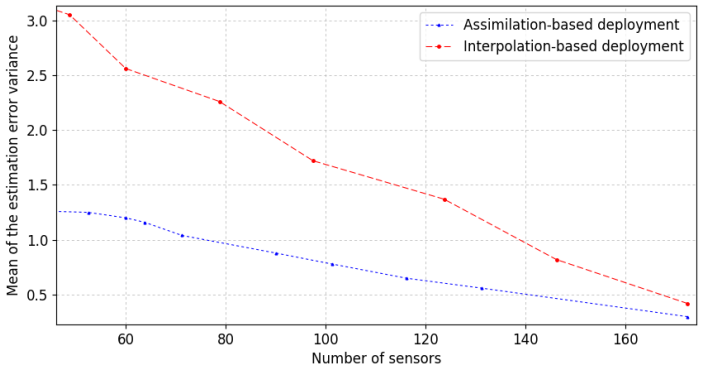


Figure 6.3: Comparison between data-assimilation-based and interpolation-based deployments.

6.6.4 Evaluation of the coverage results

We now evaluate the optimal coverage results of our deployment model while analyzing the impact of sensing errors. Results are depicted in Fig. 6.4 where the overall

deployment cost is evaluated in function of both the assimilation error and the sensing error of sensors. We notice that the maximum improvement of the assimilation depends on the quality of sensors. Indeed, when sensors are not perfect, the least assimilation error that we can get is equal to $1\mu g^2/m^6$. We also notice that a minimum number of sensors is required in order to be able to use the assimilation technique. This minimum number is equal to 50 sensors in our simulations. This means that in order to reduce the error of the physical model, we need 50 sensors or more. Finally, Fig. 6.4 shows that the more the tolerated assimilation error, the less the impact of the quality of sensors.

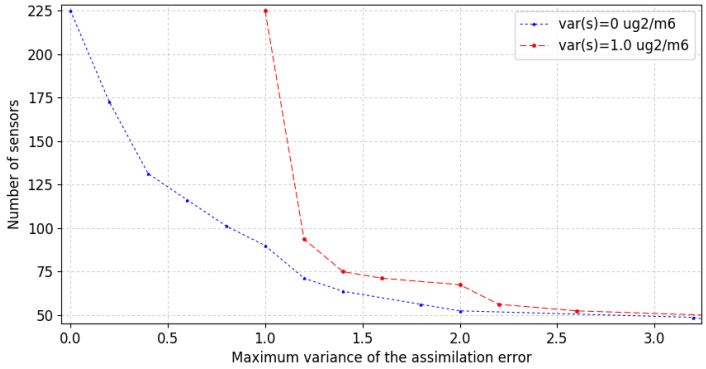


Figure 6.4: Deployment results depending on sensing errors.

6.6.5 Evaluation of the connectivity results

Finally, we evaluate the impact of the connectivity technology on the use of the deployment budget and the quality of data assimilation. We consider two different types of nodes depending on their communication capabilities: nodes with a communication range equal to 100m that we consider as short range communication nodes (like 802.15.4 for instance); and nodes with a communication range equal to 500m that we consider as long-range communication nodes (like LoRa for instance). We vary the deployment budget and depict the resulting assimilation error in Fig. 6.5.

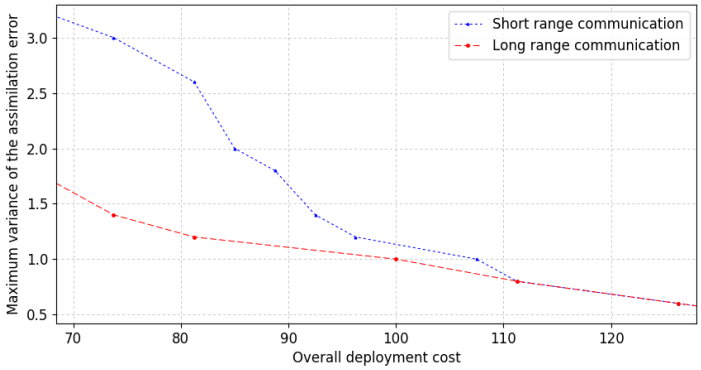


Figure 6.5: Impact of the communication technology on the deployment results.

Results show that using long range communications leads to less assimilation error when compared to short range communications with respect to the same value of the deployment budget. This is explained by the relay nodes added just to ensure connectivity in the case of short range communications, which means that the deployment optimization of some nodes is performed to improve connectivity but not necessarily coverage. Fig. 6.5 also shows that when the deployment budget increases, the difference between the two communication technologies decreases because the deployment of coverage nodes becomes so dense that the network is usually already connected even when using short range communications.

6.7 Performance evaluation of the scheduling approach

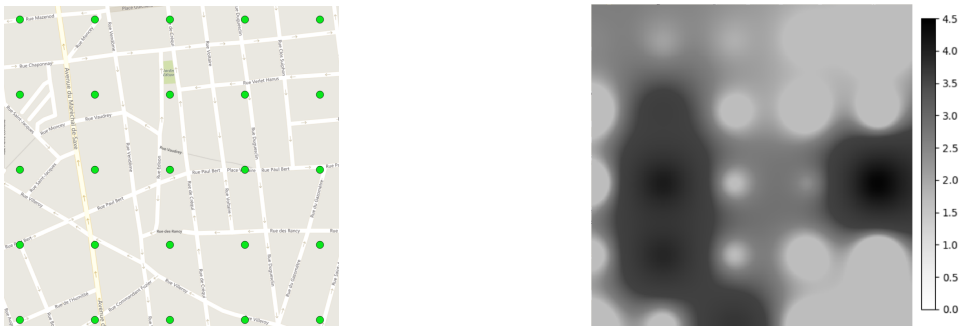
In this section, we present the simulations that we have performed in order to evaluate our scheduling approach. We first present the data set that we used and the common simulation parameters. Then, we provide a proof-of-concept and evaluate the physical models' coverage results. Finally, we assess the impact of sensing frequency and transmission power on the network lifetime.

6.7.1 Dataset

We perform the evaluation of our proposal on monthly pollution data corresponding to the 2008 Nitrogen Dioxide (NO_2) concentrations in the Lyon district of La-Part-Dieu. We consider a region of interest which has a spatial resolution of 100 meters. The monitoring region is depicted in Fig. 6.6a with a set of 25 already deployed pollution sensors. We calculate the correlation coefficients $\mathcal{W}_{pq}$ using an exponential decay function as in the evaluation part of the data-assimilation-aware deployment approach. We also assume that the variance of the errors of the physical model are linearly correlated with its concentrations. The variance map of the physical model errors corresponding to the month of June is depicted in Fig. 6.6b. In order to consider a real scheduling scenario, we use the energy consumption characteristics of our air pollution monitoring system which is presented in detail in chapter 3. The energy consumption required for sensing, transmission and reception per month as well as the default simulation parameters are depicted in table 6.3.

Parameter	Notation	Value
Number of sensors	$\mathcal{N}$	25
Communication range of sensor nodes	$\mathcal{R}$	$500m$
Sensing error	s_p^t	$0.5\mu g^2/m^6$
Battery capacity	EI_p	518.4 kJ
Sensing monthly energy consumption	ES_p^t	129.6 kJ
Transmission monthly energy consumption	ET_p^t	129.6 kJ
Reception monthly energy consumption	ER_p^t	51.84 kJ

Table 6.3: Default scheduling parameters of the physical models' correction approach.



(a) (b)

Figure 6.6: WSN nodes' locations and simulation errors' variance ($\mu g^2/m^6$) corresponding to the district of La-Part-Dieu during June 2008.

6.7.2 Proof-of-concept: application to the La-Part-Dieu district

We run our model while considering the default simulation parameters and we set coverage requirements to $1\mu g^2/m^6$. We get an optimal network lifetime equal to 3 months as depicted in Fig. 6.7 (from left to right: January, February and March). We notice that the number of active sensors is different depending on the time period. In fact, the correction of the air pollution physical model requires more nodes depending on the variability of air pollution within each time period. We also notice that some nodes are active in more than one time period. Indeed, according to the default simulation values, the battery capacity allows sensor nodes to run up to two months without stop.

We also evaluate the assimilation error provided by the sensor nodes that are active during the months of January, February and March. Results are depicted in the same Fig. 6.7. We notice that the assimilation error never exceeds the coverage requirement set to $1\mu g^2/m^6$ thanks to the coverage constraints of our optimization model.

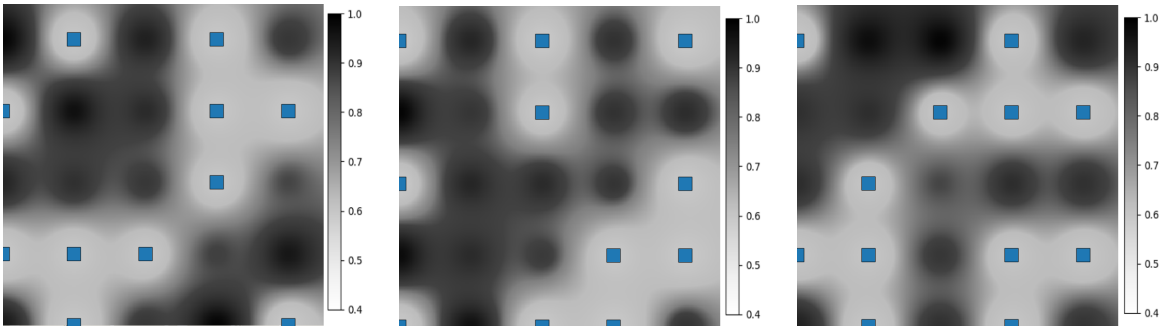


Figure 6.7: Proof-of-concept of the scheduling model: optimal WSN activity scheduling and the corresponding estimation errors' variance ($\mu g^2/m^6$) while considering coverage requirements equal to $1\mu g^2/m^6$. Active sensors are depicted in blue squares.

6.7.3 Impact of coverage requirements on the network lifetime

In this simulation scenario, we investigate the impact of coverage requirements on the maximization of the network lifetime. We consider 3 different values of the battery capacity and we vary the coverage requirements from $0.5\mu g^2/m^6$ to $2.5\mu g^2/m^6$. The

obtained results are depicted in Fig. 6.8. We notice that the higher the assimilation error threshold, the higher the network lifetime. This is expected since with less coverage requirements, less sensors are activated during each time period allowing us to get multiple subsets of sensor nodes capable, each of which, of ensuring coverage and connectivity constraints.

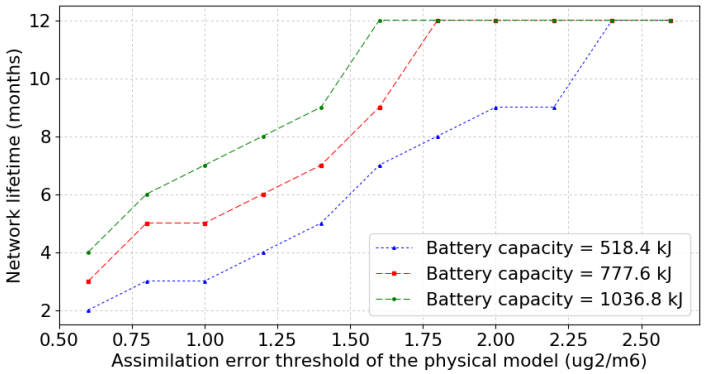


Figure 6.8: Impact of battery capacity on the scheduling results.

In terms of battery capacity impact, as expected the higher the initial amount of energy per node, the largest the network lifetime. However, doubling the battery capacity does not always involve twice lifetime as shown in the curves when the assimilation error threshold is equal to 1.4 $\mu g^2/m^6$. Indeed, this depends on the variability of pollution within each time period and which may require different number of active nodes.

6.7.4 Impact of sensing frequency on the network lifetime

We now evaluate the impact of sensing quality on the network lifetime. We consider two scenarios while varying the sensing frequency of nodes: in the first case, sensors analyze the quality of the air without stop whereas in the second case, sensing is performed during only 30 seconds every minute. It is worth mentioning that in the latter, the node is kept active during the non sensing 30 seconds which is necessary for the air pollution electrochemical sensing probes. Reducing the sensing frequency impacts both energy consumption and the correction of the physical model. In order to understand how the sensing frequency impacts the network lifetime, we depict in Fig. 6.9 the obtained results while varying the coverage requirements. Results show that first, coverage requirements cannot be met with low sensing frequency for an assimilation error threshold that is less than $1\mu g^2/m^6$. Indeed, with low sensing frequency, the sensing error of nodes goes up from $0.5\mu g^2/m^6$ to $1\mu g^2/m^6$. Moreover, we still get lower network lifetime when using low sensing frequency. However, as the assimilation error threshold of the physical model goes up, we tend to get the same results for both sensing techniques because the sensing errors become tolerable with respect to the assimilation process.

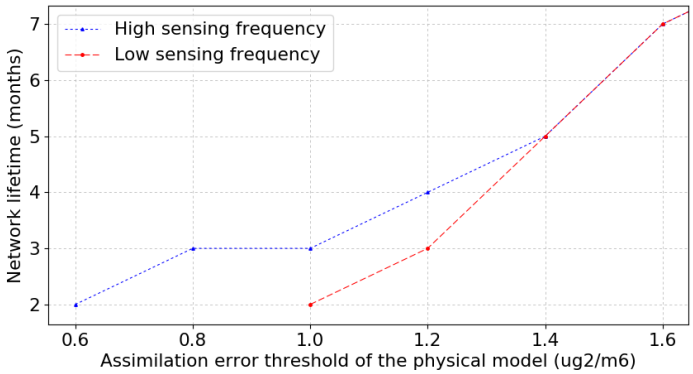


Figure 6.9: Impact of sensing quality on the scheduling results.

6.7.5 Impact of transmission power on the network lifetime

In the last simulation scenario, we analyze the impact of transmission power on the network lifetime. Increasing the transmission power allows us to get larger transmission range but at the cost of energy consumption. We consider three different cases: i) high transmission power (our LoRa powered nodes with 20dbm transmission power) allowing us to get one-hop communication to the sink node (communication range equal to 500m); ii) low transmission power (14 dbm) with a communication range equal to 200 in non-line of sight while considering the sink node in the corner of the map; and finally iii) low transmission power with the sink node in the center. Results are depicted in Fig. 6.9 and show that despite the multi-hop communication, low transmission power leads to better network lifetime compared to high transmission power. However, this is not always the case as the results also show that one-hop communication is preferable over multi-hop communication if the sink node is not well positioned (in the corner rather than in the center of the map for instance).

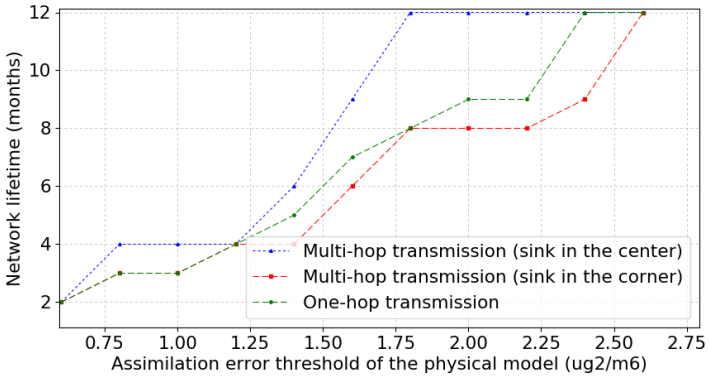


Figure 6.10: Impact of transmission power on the scheduling results.

6.8 Conclusion

In this chapter, we focused on the effective data assimilation of air pollution measurements in order to correct the physical models' simulations. We tackled first the deployment issue of sensor networks and proposed a mixed integer programming model and a heuristic algorithm allowing to ensure both coverage and connectivity. We then focused on extending the lifetime of the network by scheduling the sensing activity of sensors. We designed a mathematical model and a heuristic algorithm to determine which sensors should be turned off based on the collected data in order to let the network operate longer while ensuring the application requirements. We applied our deployment and scheduling approaches on a dataset of the Lyon City. We have shown that the assimilation-based deployment outperforms the interpolation-based one. We have also assessed by simulation the impact of the quality of sensors and their communication range on the deployment and scheduling results.

Chapter 7

Conclusions and Future Works

7.1 Main conclusions

In this thesis, we focused on the use of wireless sensor networks for three main applications of air pollution monitoring: air quality mapping, air pollution threshold crossings' detection and the correction of physical models' simulations. We addressed the deployment and scheduling issues of wireless sensor networks and presented several optimization approaches ensuring the monitoring of air pollution and the sensor network connectivity with the minimum financial cost while maximizing the lifetime of the network. Unlike the inadequate related works, which are either generic or rely on a detection model, we take into account the nature of the phenomenon in order to provide realistic optimization models and heuristic algorithms. Indeed, in our proposed approaches, we use existing pollution data and the available simulations of physical models in order to incorporate the characteristics of the air pollution phenomenon in the design of WSN-based pollution monitoring systems.

In order to study the effectiveness of our deployment and scheduling methods, we evaluated the impact of the parameters of our solutions on the obtained deployment and scheduling results while using real data sets mainly from Lyon and London cities. Among our conclusions, we highlight the fact that sensors should be placed close to the effective height of pollution sources and the fact that the size of the resulting network depends on the degree of the variations of pollution concentrations within the deployment region. We also concluded that the desired monitoring precision impacts well the density of sensors and hence the connectivity of the network.

Besides our theoretical proposals regarding the WSN deployment and scheduling techniques, we designed from scratch an energy-efficient and cost-effective air pollution monitoring platform. Compared to existing platforms, we leveraged the LoRa communication technology and the use of low-power microcontrollers and sensing probes in order to reduce the energy consumption of our measurement nodes. The analysis of our first results allowed us to show that our nodes can provide accurate good measurements compared to some reference sensors that are available on the market. The obtained dataset also helped us study the impact of weather conditions on communication links especially when sensors are powered using solar panels.

7.2 Extensions and future works

Both our theoretical works and our monitoring platform can be extended in order to provide a better monitoring of air quality and many other environmental applications.

Future extensions of our optimization solutions

In addition to the sensor networks' connectivity, our main focus in this thesis was on appropriate coverage formulations of air pollution monitoring while considering different application cases. Motivated by the fact that pollution sensors operate well when they are static [136], we designed our optimization models for mainly static sensor nodes. However, sinks can be considered as mobile nodes, and this can be integrated in our models based on existing mobile-sinks' formulations such as the work of [15]. The idea is to first consider a set of timeframes which characterize the mobility of sinks nodes; i.e. a mobile sink can change its positions in each timeframe. In order to ensure that sensors are connected to mobile sinks, the flow concept is again used and should be formulated in order to ensure that the flow is conservative in the network in each timeframe. This means that in each timeframe, each sensor generates a flow unit and the mobile sink nodes receive all the generated units.

In addition to mobility, we are working on extending our current formulations in order to take into account the impact of the different urban parameters, such as the structure of streets. Our coverage formulations can be also extended in order to ensure the monitoring of multiple pollutants at the same time by leveraging the effects of cross-sensitivity of gas sensing probes.

Although our optimization solutions are designed mainly for air pollution monitoring, they can also work for other environmental applications. We are therefore planning to study the performance of our solutions when dealing with different environmental applications.

Future extensions of our monitoring platform

The extensive simulations that we performed on our deployment and scheduling approaches allowed us to show their effectiveness. However, we still need to validate their performance using real deployments. Our monitoring platform can in fact help us in this purpose thanks to the promising first results that we have got from our deployment case study in the Lyon city. Besides, our lab-designed sensors are not limited to static monitoring and can be adapted to different sensing paradigms like being deployed on top of vehicles, attached to drones or carried by users for crowdsensing. However, some extensions need to be performed before using our nodes in those monitoring scenarios. First, our data processing techniques need to be adapted to the mobile case by filtering out the effects of the airflow on the sensing probes during mobility. Then, we might need to upgrade the current microcontroller in order to consider more computation tasks, which will come of course at the cost of power consumption. Finally, for crowdsensing applications, we need to make the box of the nodes more resistant to damages and more user friendly.

Bibliography

[1] World Health Organization, "Burden of disease from the joint effects of household and ambient air pollution for 2016." On http://www.who.int/airpollution/data/AP_joint_effect_BoD_results_May2018.pdf?ua=1 [access: 2018-09-06].

[2] World Health Organization, "The impact of the environment on children's health." On http://apps.who.int/iris/bitstream/handle/10665/254678/WHO-FWC-IHE-17.01-eng.pdf [access: 2018-09-06].

[3] C. A. Pope III, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, and G. D. Thurston, "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution," *Jama*, vol. 287, no. 9, pp. 1132–1141, 2002.

[4] G. Bowatte, C. Lodge, A. J. Lowe, B. Erbas, J. Perret, M. J. Abramson, M. Matheson, and S. C. Dharmage, "The influence of childhood traffic-related air pollution exposure on asthma, allergy and sensitization: a systematic review and a meta-analysis of birth cohort studies," *Allergy*, vol. 70, no. 3, pp. 245–256, 2015.

[5] Environmental Protection Agency, "Air pollution: current and future challenges." Available on https://www.epa.gov/clean-air-act-overview/air-pollution-current-and-future-challenges, [access: 2018-09-01].

[6] B. Beckerman, M. Jerrett, J. R. Brook, D. K. Verma, M. A. Arain, and M. M. Finkelstein, "Correlation of nitrogen dioxide with other traffic pollutants near a major expressway," *Atmospheric Environment*, vol. 42, no. 2, pp. 275–290, 2008.

[7] J. C. von Fischer, D. Cooley, S. Chamberlain, A. Gaylord, C. J. Griebenow, S. P. Hamburg, J. Salo, R. Schumacher, D. Theobald, and J. Ham, "Rapid, vehicle-based identification of location and magnitude of urban natural gas pipeline leaks," *Environmental Science & Technology*, vol. 51, no. 7, pp. 4091–4099, 2017.

[8] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele, "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervasive and Mobile Computing*, vol. 16, pp. 268–285, 2015.

[9] S. O. Ahmed, R. Mazloum, and H. Abou-Ali, "Spatiotemporal interpolation of air pollutants in the greater cairo and the delta, egypt," *Environmental research*, vol. 160, pp. 27–34, 2018.

[10] J. Evans, A. van Donkelaar, R. V. Martin, R. Burnett, D. G. Rainham, N. J. Birkett, and D. Krewski, "Estimates of global mortality attributable to particulate air pollution using satellite imagery," *Environmental research*, vol. 120, pp. 33–42, 2013.

[11] M. Jerrett, A. Arain, P. Kanaroglou, B. Beckerman, D. Potoglou, T. Sahsuvaroglu, J. Morrison, and C. Giovis, "A review and evaluation of intraurban air pollution exposure models," *Journal of Exposure Science and Environmental Epidemiology*, vol. 15, no. 2, pp. 185–204, 2005.

[12] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer networks*, vol. 38, no. 4, pp. 393–422, 2002.

[13] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[14] S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D. M. Broday, and B. Fishbain, "On the feasibility of measuring urban air pollution by wireless distributed sensor networks," *Science of The Total Environment*, vol. 502, pp. 537–547, 2015.

[15] M. E. Keskin, İ. K. Altınel, N. Aras, and C. Ersoy, "Wireless sensor network lifetime maximization by optimal sensor deployment, activity scheduling, data routing and sink mobility," *Ad Hoc Networks*, vol. 17, pp. 18–36, 2014.

[16] O. Saukh, D. Hasenfratz, and L. Thiele, "Route selection for mobile sensor nodes on public transport networks," *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, no. 3, pp. 307–321, 2014.

[17] Y. Yang, Z. Zheng, K. Bian, L. Song, and Z. Han, "Real-time profiling of fine-grained air quality index distribution using uav sensing," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 186–198, 2018.

[18] A. Anjomshoaa, S. Mora, P. Schmitt, and C. Ratti, "Challenges of drive-by iot sensing for smart cities: City scanner case study," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pp. 1112–1120, ACM, 2018.

[19] B. Maag, O. Saukh, D. Hasenfratz, and L. Thiele, "Pre-deployment testing, augmentation and calibration of cross-sensitive sensors.," in *EWSN*, pp. 169–180, 2016.

[20] B. Maag, Z. Zhou, O. Saukh, and L. Thiele, "Scan: Multi-hop calibration for mobile sensor arrays," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, p. 19, 2017.

[21] S. De Vito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones, and G. Di Francia, "Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches," *Sensors and Actuators B: Chemical*, vol. 255, pp. 1191–1210, 2018.

[22] W. Yi, K. Lo, T. Mak, K. Leung, Y. Leung, and M. Meng, "A survey of wireless sensor network based air pollution monitoring systems," *Sensors*, vol. 15, no. 12, pp. 31392–31427, 2015.

[23] K. Chakrabarty, S. S. Iyengar, H. Qi, and E. Cho, "Grid coverage for surveillance and target location in distributed sensor networks," *Computers, IEEE Transactions on*, vol. 51, no. 12, pp. 1448–1453, 2002.

[24] M. Rebai, H. Snoussi, F. Hnaien, L. Khoukhi, *et al.*, "Sensor deployment optimization methods to achieve both coverage and connectivity in wireless sensor networks," *Computers & Operations Research*, vol. 59, pp. 11–21, 2015.

[25] S. Sengupta, S. Das, M. Nasir, and B. K. Panigrahi, "Multi-objective node deployment in wsns: In search of an optimal trade-off among coverage, lifetime, energy consumption, and connectivity," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 1, pp. 405–416, 2013.

[26] F. Tashtarian, M. H. Y. Moghaddam, K. Sohraby, and S. Effati, "On maximizing the lifetime of wireless sensor networks in event-driven applications with mobile sinks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 3177–3189, 2015.

[27] H. Yetgin, K. T. K. Cheung, M. El-Hajjar, and L. H. Hanzo, "A survey of network lifetime maximization techniques in wireless sensor networks," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 828–854, 2017.

[28] S. Mini, S. K. Udgata, and S. L. Sabat, "Sensor deployment and scheduling for target coverage problem in wireless sensor networks," *IEEE Sensors Journal*, vol. 14, no. 3, pp. 636–644, 2014.

[29] A. Boubrima, W. Bechkit, and H. Rivano, "Optimal wsn deployment models for air pollution monitoring," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 2723–2735, 2017.

[30] A. Boubrima, W. Bechkit, and H. Rivano, "On the deployment of wireless sensor networks for air quality mapping: Optimization models and algorithms," *IEEE/ACM Transactions on Networking*, vol. 27, pp. 1629–1642, Aug 2019.

[31] A. Boubrima, W. Bechkit, and H. Rivano, "On the optimization of wsn deployment for sensing physical phenomena: Applications to urban air pollution monitoring," in *The Philosophy of Mission-Oriented Wireless Sensor Networks*, Springer, 2019.

[32] A. Boubrima, A. Boukerche, W. Bechkit, and H. Rivano, "Wsn scheduling for energy-efficient correction of environmental modelling," in *IEEE MASS 2018-15th IEEE International Conference on Mobile Ad-hoc and Sensor Systems*, pp. 1–8, 2018.

[33] A. Boubrima, W. Bechkit, H. Rivano, and L. Soulhac, "Leveraging the potential of wsn for an efficient correction of air pollution fine-grained simulations," in

Computer Communication and Networks (ICCCN), 2018 27th International Conference on, IEEE, 2018.

[34] A. Boubrima, W. Bechkit, and H. Rivano, "A new wsn deployment approach for air pollution monitoring," in *Consumer Communications & Networking Conference (CCNC), 2017 14th IEEE Annual*, pp. 455–460, IEEE, 2017.

[35] A. Boubrima, W. Bechkit, and H. Rivano, "Error-bounded air quality mapping using wireless sensor networks," in *Local Computer Networks (LCN), 2016 IEEE 41st Conference on*, pp. 380–388, IEEE, 2016.

[36] A. Boubrima, W. Bechkit, and H. Rivano, "Optimal deployment of dense wsn for error bounded air pollution mapping," in *Distributed Computing in Sensor Systems (DCOSS), 2016 International Conference on*, pp. 102–104, IEEE, 2016.

[37] A. Boubrima, F. Matigot, W. Bechkit, H. Rivano, and A. Ruas, "Optimal deployment of wireless sensor networks for air pollution monitoring," in *Computer Communication and Networks (ICCCN), 2015 24th International Conference on*, pp. 1–7, IEEE, 2015.

[38] A. Boubrima, W. Bechkit, H. Rivano, and A. Ruas, "Wireless sensor networks deployment for air pollution monitoring," in *TAP 2016-21st International Transport and Air Pollution Conference*, 2016.

[39] A. Boubrima, W. Bechkit, H. Rivano, and L. Soulhac, "Cost-precision tradeoffs in 3d air pollution mapping using wsn," in *Advances in Ubiquitous Networking 2*, pp. 191–203, Springer, 2017.

[40] A. Boubrima, W. Bechkit, H. Rivano, and L. Soulhac, "Poster: Toward a better monitoring of air pollution using mobile wireless sensor networks," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, pp. 534–536, ACM, 2017.

[41] M. Younis and K. Akkaya, "Strategies and techniques for node placement in wireless sensor networks: A survey," *Ad Hoc Networks*, vol. 6, no. 4, pp. 621–655, 2008.

[42] A. Ghosh and S. K. Das, "Coverage and connectivity issues in wireless sensor networks," *Mobile, wireless, and sensor networks: Technology, applications, and future directions*, pp. 221–256, 2006.

[43] M. R. Senouci, A. Mellouk, K. Asnoune, and F. Y. Bouhidel, "Movement-assisted sensor deployment algorithms: A survey and taxonomy," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2493–2510, 2015.

[44] I. Khoufi, P. Minet, A. Laouiti, and S. Mahfoudh, "Survey of deployment algorithms in wireless sensor networks: coverage and connectivity issues and challenges," *International Journal of Autonomous and Adaptive Communications Systems (IJAACS)*, vol. 10, no. 4, pp. 341–390, 2017.

[45] N. A. A. Aziz, K. A. Aziz, and W. Z. W. Ismail, "Coverage strategies for wireless sensor networks," *World academy of science, Engineering and technology*, vol. 50, pp. 145–150, 2009.

[46] C. Zhu, C. Zheng, L. Shu, and G. Han, "A survey on coverage and connectivity issues in wireless sensor networks," *Journal of Network and Computer Applications*, vol. 35, no. 2, pp. 619–632, 2012.

[47] A. K. Idrees, K. Deschinkel, M. Salomon, and R. Couturier, "Perimeter-based coverage optimization to improve lifetime in wireless sensor networks," *Engineering Optimization*, vol. 48, no. 11, pp. 1951–1972, 2016.

[48] B. Wang, "Coverage problems in sensor networks: A survey," *ACM Computing Surveys (CSUR)*, vol. 43, no. 4, p. 32, 2011.

[49] S. Biswas, R. Das, and P. Chatterjee, "Energy-efficient connected target coverage in multi-hop wireless sensor networks," in *Industry Interactive Innovations in Science, Engineering and Technology*, pp. 411–421, Springer, 2018.

[50] J. Lin, W. Xiao, F. L. Lewis, and L. Xie, "Energy-efficient distributed adaptive multisensor scheduling for target tracking in wireless sensor networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 6, pp. 1886–1896, 2009.

[51] B. Wang, H. B. Lim, and D. Ma, "A survey of movement strategies for improving network coverage in wireless sensor networks," *Computer Communications*, vol. 32, no. 13-14, pp. 1427–1436, 2009.

[52] Y. Hao and R. Foster, "Wireless body sensor networks for health-monitoring applications," *Physiological measurement*, vol. 29, no. 11, p. R27, 2008.

[53] A. Ghosh and S. K. Das, "Coverage and connectivity issues in wireless sensor networks: A survey," *Pervasive and Mobile Computing*, vol. 4, no. 3, pp. 303–334, 2008.

[54] M. K. Watfa, H. Al-Hassanieh, and S. Salmen, "A novel solution to the energy hole problem in sensor networks," *Journal of network and computer applications*, vol. 36, no. 2, pp. 949–958, 2013.

[55] A. Bachir, W. Bechkit, Y. Challal, and A. Bouabdallah, "Joint connectivity-coverage temperature-aware algorithms for wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 7, pp. 1923–1936, 2015.

[56] J. Chen, E. Shen, and Y. Sun, "The deployment algorithms in wireless sensor net works: A survey," *Information Technology Journal*, vol. 8, no. 3, pp. 293–301, 2009.

[57] S. Oktug, A. Khalilov, and H. Tezcan, "3d coverage analysis under heterogeneous deployment strategies in wireless sensor networks," in *Telecommunications, 2008. AICT'08. Fourth Advanced International Conference on*, pp. 199–204, IEEE, 2008.

[58] B. Liu, O. Dousse, P. Nain, and D. Towsley, "Dynamic coverage of mobile sensor networks," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 2, pp. 301–311, 2013.

[59] D. S. Deif and Y. Gadallah, "Classification of wireless sensor networks deployment techniques," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 834–855, 2014.

[60] Y. Wang, Y. Zhang, J. Liu, and R. Bhandari, "Coverage, connectivity, and deployment in wireless sensor networks," in *Recent Development in Wireless Sensor and Ad-hoc Networks*, pp. 25–44, Springer, 2015.

[61] K. Chakrabarty, S. S. Iyengar, H. Qi, and E. Cho, "Coding theory framework for target location in distributed sensor networks," in *Information Technology: Coding and Computing, 2001. Proceedings. International Conference on*, pp. 130–134, IEEE, 2001.

[62] F. Glover, "Improved linear integer programming formulations of nonlinear integer problems," *Management Science*, vol. 22, no. 4, pp. 455–460, 1975.

[63] S. Meguerdichian and M. Potkonjak, "Low power 0/1 coverage and scheduling techniques in sensor networks," tech. rep., UCLA Technical Reports 030001, 2003.

[64] V. Chvatal, "A greedy heuristic for the set-covering problem," *Mathematics of operations research*, vol. 4, no. 3, pp. 233–235, 1979.

[65] İ. K. Altınel, N. Aras, E. Güney, and C. Ersoy, "Binary integer programming formulation and heuristics for differentiated coverage in heterogeneous sensor networks," *Computer Networks*, vol. 52, no. 12, pp. 2419–2431, 2008.

[66] I. K. Altinel, N. Aras, E. Guney, and C. Ersoy, "Effective coverage in sensor networks: binary integer programming formulations and heuristics," in *ICC'06. IEEE International Conference on*, vol. 9, pp. 4014–4019, IEEE, 2006.

[67] X. Wang, G. Xing, Y. Zhang, C. Lu, R. Pless, and C. Gill, "Integrated coverage and connectivity configuration in wireless sensor networks," in *Proceedings of the 1st international conference on Embedded networked sensor systems*, pp. 28–39, ACM, 2003.

[68] S. S. Dhillon and K. Chakrabarty, "Sensor placement for effective coverage and surveillance in distributed sensor networks," in *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE*, vol. 3, pp. 1609–1614, IEEE, 2003.

[69] X. Chang, R. Tan, G. Xing, Z. Yuan, C. Lu, Y. Chen, and Y. Yang, "Sensor placement algorithms for fusion-based surveillance networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 8, pp. 1407–1414, 2011.

[70] Z. Yuan, R. Tan, G. Xing, C. Lu, Y. Chen, and J. Wang, "Fast sensor placement algorithms for fusion-based target detection," in *Real-Time Systems Symposium, 2008*, pp. 103–112, IEEE, 2008.

[71] Y. Zou and K. Chakrabarty, "Sensor deployment and target localization based on virtual forces," in *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 2, pp. 1293–1303, IEEE, 2003.

[72] Y. Zou and K. Chakrabarty, "Sensor deployment and target localization in distributed sensor networks," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 3, no. 1, pp. 61–91, 2004.

[73] J. Chen, S. Li, and Y. Sun, "Novel deployment schemes for mobile sensor networks," *Sensors*, vol. 7, no. 11, pp. 2907–2919, 2007.

[74] K. Akkaya and M. Younis, "Coverage-aware and connectivity-constrained actor positioning in wireless sensor and actor networks," in *26th IEEE International Performance Computing and Communications Conference (IPCCC 2007)*.

[75] Z. Yong and W. Li, "A sensor deployment algorithm for mobile wireless sensor networks," in *Control and Decision Conference, 2009. CCDC'09. Chinese*, pp. 4606–4611, IEEE, 2009.

[76] A. Howard, M. J. Matarić, and G. S. Sukhatme, "Mobile sensor network deployment using potential fields: A distributed, scalable solution to the area coverage problem," in *Distributed Autonomous Robotic Systems 5*, pp. 299–308, Springer, 2002.

[77] S. Li, C. Xu, W. Pan, and Y. Pan, "Sensor deployment optimization for detecting maneuvering targets," in *Information Fusion, 2005 8th International Conference on*, vol. 2, pp. 7–pp, IEEE, 2005.

[78] Y. Zou and K. Chakrabarty, "A distributed coverage-and connectivity-centric technique for selecting active nodes in wireless sensor networks," *IEEE Transactions on Computers*, vol. 54, no. 8, pp. 978–991, 2005.

[79] X.-M. Hu, J. Zhang, Y. Yu, H. S.-H. Chung, Y.-L. Li, Y.-H. Shi, and X.-N. Luo, "Hybrid genetic algorithm using a forward encoding scheme for lifetime maximization of wireless sensor networks," *IEEE transactions on evolutionary computation*, vol. 14, no. 5, pp. 766–781, 2010.

[80] Y. Lin, J. Zhang, H. S.-H. Chung, W. H. Ip, Y. Li, and Y.-H. Shi, "An ant colony optimization approach for maximizing the lifetime of heterogeneous wireless sensor networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 3, pp. 408–420, 2012.

[81] J. Du, K. Wang, H. Liu, and D. Guo, "Maximizing the lifetime of *k*-discrete barrier coverage using mobile sensors," *IEEE Sensors Journal*, vol. 13, no. 12, pp. 4690–4701, 2013.

[82] Y. B. Turkogullari, N. Aras, I. K. Altinel, and C. Ersoy, "Optimal placement and activity scheduling to maximize coverage lifetime in wireless sensor networks," in *Computer and information sciences, 2007. iscis 2007. 22nd international symposium on*, pp. 1–6, IEEE, 2007.

[83] C.-P. Chen, S. C. Mukhopadhyay, C.-L. Chuang, M.-Y. Liu, and J.-A. Jiang, "Efficient coverage and connectivity preservation with load balance for wireless sensor networks," *IEEE sensors journal*, vol. 15, no. 1, pp. 48–62, 2015.

[84] Z. Lu, W. W. Li, and M. Pan, "Maximum lifetime scheduling for target coverage and data collection in wireless sensor networks," *IEEE Transactions on vehicular technology*, vol. 64, no. 2, pp. 714–727, 2015.

[85] V. Roy, A. Simonetto, and G. Leus, "Spatio-temporal sensor management for environmental field estimation," *Signal Processing*, vol. 128, pp. 369–381, 2016.

[86] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies," *Journal of Machine Learning Research*, vol. 9, no. Feb, pp. 235–284, 2008.

[87] P. G. Liaskovitis and C. Schurgers, "Leveraging redundancy in sampling-interpolation applications for sensor networks: A spectral approach," *ACM Transactions on Sensor Networks (TOSN)*, vol. 7, no. 2, p. 12, 2010.

[88] J. Hao, B. Zhang, Z. Jiao, and S. Mao, "Adaptive compressive sensing based sample scheduling mechanism for wireless sensor networks," *Pervasive and Mobile Computing*, vol. 22, pp. 113–125, 2015.

[89] X. Wu and M. Liu, "In-situ soil moisture sensing: measurement scheduling and estimation using compressive sensing," in *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, pp. 1–12, ACM, 2012.

[90] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg, "Robust sensor placements at informative and communication-efficient locations," *ACM Transactions on Sensor Networks (TOSN)*, vol. 7, no. 4, p. 31, 2011.

[91] J. Ranieri, A. Chebira, and M. Vetterli, "Near-optimal sensor placement for linear inverse problems," *IEEE Transactions on signal processing*, vol. 62, no. 5, pp. 1135–1146, 2014.

[92] X. Deng, B. Wang, W. Liu, and L. T. Yang, "Sensor scheduling for multi-modal confident information coverage in sensor networks.," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 3, pp. 902–913, 2015.

[93] W. Du, Z. Xing, M. Li, B. He, L. H. C. Chua, and H. Miao, "Sensor placement and measurement of wind for water quality studies in urban reservoirs," *ACM Transactions on Sensor Networks (TOSN)*, vol. 11, no. 3, p. 41, 2015.

[94] W. Du, Z. Xing, M. Li, B. He, L. H. C. Chua, and H. Miao, "Optimal sensor placement and measurement of wind for water quality studies in urban reservoirs," in *Proceedings of the 13th international symposium on Information processing in sensor networks*, pp. 167–178, IEEE Press, 2014.

[95] X. Wang, X. Wang, G. Xing, J. Chen, C.-X. Lin, and Y. Chen, "Intelligent sensor placement for hot server detection in data centers," *IEEE Transactions on parallel and distributed systems*, vol. 24, no. 8, pp. 1577–1588, 2013.

[96] X. Wang, X. Wang, G. Xing, J. Chen, C.-X. Lin, and Y. Chen, "Towards optimal sensor placement for hot server detection in data centers," in *Distributed Computing Systems (ICDCS), 2011 31st International Conference on*, pp. 899–908, IEEE, 2011.

[97] Z. Khalfallah, I. Fajjariy, N. Aitsaadiz, R. Langar, and G. Pujolle, "A new wsn deployment algorithm for water pollution monitoring in amazon rainforest rivers," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, pp. 267–273, IEEE, 2013.

[98] Z. Khalfallah, I. Fajjari, N. Aitsaadi, P. Rubin, and G. Pujolle, "A novel 3d underwater wsn deployment strategy for full-coverage and connectivity in rivers," in *Communications (ICC), 2016 IEEE International Conference on*, pp. 1–7, IEEE, 2016.

[99] Z. Khalfallah, I. Fajjariz, N. Aitsaadiz, R. Langar, and G. Pujolle, "2d-ubda: A novel 2-dimensional underwater wsn barrier deployment algorithm," in *IFIP Networking Conference (IFIP Networking), 2015*, pp. 1–8, IEEE, 2015.

[100] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, and L. Thiele, "Pushing the spatio-temporal resolution limit of urban air pollution maps," in *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, pp. 69–77, IEEE, 2014.

[101] J. S. Apte, K. P. Messier, S. Gani, M. Brauer, T. W. Kirchstetter, M. M. Lunden, J. D. Marshall, C. J. Portier, R. C. Vermeulen, and S. P. Hamburg, "High-resolution air pollution mapping with google street view cars: exploiting big data," *Environmental Science & Technology*, vol. 51, no. 12, pp. 6999–7008, 2017.

[102] A. Anjomshoaa, F. Duarte, D. Rennings, T. Matarazzo, P. de Souza, and C. Ratti, "City scanner: Building and scheduling a mobile sensing platform for smart city services," *IEEE Internet of Things Journal*, 2018.

[103] Y. Yang, Z. Bai, Z. Hu, Z. Zheng, K. Bian, and L. Song, "Aqnet: Fine-grained 3d spatio-temporal air quality monitoring by aerial-ground wsn," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–2, IEEE, 2018.

[104] R. Petrolo, Y. Lin, and E. Knightly, "Astro: Autonomous, sensing, and tetherless networked drones," in *Proceedings of the 4th ACM Workshop on Micro Aerial Vehicle Networks, Systems, and Applications*, pp. 1–6, ACM, 2018.

[105] B. Maag, Z. Zhou, and L. Thiele, "W-air: Enabling personal air pollution monitoring on wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, p. 24, 2018.

[106] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G.-S. Ahn, and A. T. Campbell, "Bikenet: A mobile sensing system for cyclist experience mapping," *ACM Transactions on Sensor Networks (TOSN)*, vol. 6, no. 1, p. 6, 2009.

[107] Y. Zhuang, F. Lin, E.-H. Yoo, and W. Xu, "Airsense: A portable context-sensing device for personal air quality monitoring," in *Proceedings of the 2015 Workshop on Pervasive Wireless Healthcare*, pp. 17–22, ACM, 2015.

[108] P. Dutta, P. M. Aoki, N. Kumar, A. Mainwaring, C. Myers, W. Willett, and A. Woodruff, "Common sense: participatory urban sensing using a network of handheld air quality monitors," in *Proceedings of the 7th ACM conference on embedded networked sensor systems*, pp. 349–350, ACM, 2009.

[109] N. Nikzad, N. Verma, C. Ziftci, E. Bales, N. Quick, P. Zappi, K. Patrick, S. Dasgupta, I. Krueger, T. Š. Rosing, *et al.*, "Citisense: Improving geospatial environmental assessment of air quality using a wireless personal exposure monitoring system," in *Proceedings of the conference on Wireless Health*, p. 11, ACM, 2012.

[110] AirMontech Consortium and others, "Airmontech newsletter," 2013.

[111] N. Hodges, S. van den Elshout, H.-J. Heich, C. Lad, K. Léger, and F. Nussio, "Citeair–common information to european air," in *Proceedings of the 19th International Conference Informatics for Environmental Protection-EnviroInfo: Networking Environmental Information, Part*, vol. 2, pp. 7–9, 2005.

[112] M. Mead, O. Popoola, G. Stewart, P. Landshoff, M. Calleja, M. Hayes, J. Baldovi, M. McLeod, T. Hodgson, J. Dicks, *et al.*, "The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks," *Atmospheric Environment*, vol. 70, pp. 186–203, 2013.

[113] L. Morawska, P. K. Thai, X. Liu, A. Asumadu-Sakyi, G. Ayoko, A. Bartonova, A. Bedini, F. Chai, B. Christensen, M. Dunbabin, *et al.*, "Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?," *Environment international*, vol. 116, pp. 286–299, 2018.

[114] N. Afshar-Mohajer, C. Zuidema, S. Sousan, L. Hallett, M. Tatum, A. M. Rule, G. Thomas, T. M. Peters, and K. Koehler, "Evaluation of low-cost electrochemical sensors for environmental monitoring of ozone, nitrogen dioxide, and carbon monoxide," *Journal of occupational and environmental hygiene*, vol. 15, no. 2, pp. 87–98, 2018.

[115] B. Berthelot, A. B. Daoud, B. Hellio, and R. Akiki, "Cairsens no2: A miniature device dedicated to the indicative measurement of nitrogen dioxide in ambient air," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 1, p. 473, 2017.

[116] J. Petajajarvi, K. Mikhaylov, A. Roivainen, T. Hanninen, and M. Pettissalo, "On the coverage of lpwans: range evaluation and channel attenuation model for lora technology," in *ITS Telecommunications (ITST), 2015 14th International Conference on*, pp. 55–59, IEEE, 2015.

[117] L. Soulhac, P. Salizzoni, P. Mejean, D. Didier, and I. Rios, "The model sirane for atmospheric urban pollutant dispersion; part II, validation of the model on a real case study," *Atmospheric Environment*, vol. 49, pp. 320–337, 2012.

[118] V. Gallart, S. Felici-Castell, M. Delamo, A. Foster, and J. J. Perez, "Evaluation of a real, low cost, urban wsn deployment for accurate environmental monitoring," in *Mobile Adhoc and Sensor Systems (MASS), 2011 IEEE 8th International Conference on*, pp. 634–639, IEEE, 2011.

[119] D. W. Wong, L. Yuan, and S. A. Perlin, "Comparison of spatial interpolation methods for the estimation of air quality data," *Journal of Exposure Science and Environmental Epidemiology*, vol. 14, no. 5, pp. 404–415, 2004.

[120] L. Contreras and C. Ferri, "Wind-sensitive interpolation of urban air pollution forecasts," *Procedia Computer Science*, vol. 80, pp. 313–323, 2016.

[121] A. Tilloy, V. Mallet, D. Poulet, C. Pesin, and F. Brocheton, "Blue-based no2 data assimilation at urban scale," *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 4, pp. 2031–2040, 2013.

[122] A. Daly and P. Zannetti, "Air pollution modeling–an overview," *Ambient air pollution*, 2007.

[123] J. M. Stockie, "The mathematics of atmospheric dispersion modeling," *Siam Review*, vol. 53, no. 2, pp. 349–372, 2011.

[124] London DataStore, "London atmospheric emissions inventory 2010." On http://data.london.gov.uk/dataset/london-atmospheric-emissions-inventory-2010 [2018-01-15].

[125] The guardian, "London ranks among worst european cities for air pollution." On `http://www.theguardian.com/environment/2011/sep/07/london-worst-european-cities-air-pollution` [2018-02-15].

[126] Wind Finder, "Wind and weather statistics london-heathrow." On http://www.windfinder.com/windstatistics/london-heathrow [2018-01-15].

[127] Camden Data, "Camden lighting point." On http://www.camdendata.info/Pages/Home.aspx [2018-07-15].

[128] G. Morales-España, J. M. Latorre, and A. Ramos, "Tight and compact milp formulation of start-up and shut-down ramping in unit commitment," *Power Systems, IEEE Transactions on*, vol. 28, no. 2, pp. 1288–1296, 2013.

[129] G. Morales-España, J. M. Latorre, and A. Ramos, "Tight and compact milp formulation for the thermal unit commitment problem," *Power Systems, IEEE Transactions on*, vol. 28, no. 4, pp. 4897–4908, 2013.

[130] S. R. Gandham, M. Dawande, R. Prakash, and S. Venkatesan, "Energy efficient schemes for wireless sensor networks with multiple mobile base stations," in *Global telecommunications conference, 2003. GLOBECOM'03. IEEE*, vol. 1, pp. 377–381, IEEE, 2003.

[131] LGFL, "London grid for learning." On http://weather.lgfl.org.uk/ [2018-06-20].

[132] A. A. L. Kama, H. Hammadou, B. Meurisse, and C. Papaix, "Le pic de pollution a paris du 12 au 17 mars 2014," tech. rep.

[133] R. H. Reichle, "Data assimilation methods in the earth sciences," *Advances in water resources*, vol. 31, no. 11, pp. 1411–1418, 2008.

[134] M. Asch, M. Bocquet, and M. Nodet, *Data assimilation: methods, algorithms, and applications*. SIAM, 2016.

[135] W. Revelle, "An introduction to psychometric theory with applications in r," 2009.

[136] A. Velasco, R. Ferrero, F. Gandino, B. Montrucchio, and M. Rebaudengo, "A mobile and low-cost system for environmental monitoring: A case study," *Sensors*, vol. 16, no. 5, p. 710, 2016.

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : Boubrima
Prénom : Ahmed

DATE de SOUTENANCE : 12/03/2019

TITRE : Deployment and Scheduling of Wireless Sensor Networks for Air Pollution Monitoring

NATURE : Doctorat

Numéro d'ordre :

Ecole doctorale : InfoMaths
Spécialité : Informatique

RESUME :

Les réseaux de capteurs sans fil (RCSF) sont largement utilisés dans les applications environnementales où l'objectif est de détecter un phénomène physique tel que la température, l'humidité, la pollution de l'air, etc. Dans ce contexte d'application, l'utilisation de RCSF permet de comprendre les variations du phénomène et donc être en mesure de prendre des décisions appropriées concernant son impact. En raison des limitations de ses méthodes de suivi traditionnelles et de sa grande variabilité spatiale et temporelle, la pollution de l'air est considérée comme l'un des principaux phénomènes physiques qui restent à étudier et à caractériser. Dans cette thèse, nous considérons trois applications concernant l'utilisation de RCSF pour le suivi de la pollution de l'air : la cartographie en temps réel de la qualité de l'air, la détection de dépassements de seuils des polluants et la correction de modèles physiques qui simulent le phénomène de dispersion de la pollution. Toutes ces applications nécessitent de déployer et d'ordonnancer minutieusement les capteurs afin de mieux comprendre la pollution atmosphérique tout en garantissant un coût de déploiement minimal et en maximisant la durée de vie du réseau. Notre objectif est de résourdre les problèmes de déploiement et d'ordonnancement tout en tenant compte des caractéristiques spécifiques du phénomène de la pollution de l'air. Nous proposons pour chaque cas d'application une approche efficace pour le déploiement de noeuds capteurs et puits. Nous proposons également une approche d'ordonnancement adaptée au cas de la correction de modèles physiques. Nos approches d'optimisation prennent en compte la nature physique de la pollution atmosphérique et intègrent les données réelles fournies par les plateformes existantes de suivi de la qualité de l'air. Dans chacune de nos approches d'optimisation, nous utilisons la programmation linéaire en nombres entiers pour concevoir des modèles d'optimisation adaptés à la résolution de petites et moyennes instances. Pour traiter les grandes instances, nous proposons des heuristiques en utilisant des techniques de relaxation linéaire. Outre nos travaux théoriques sur le suivi de la pollution atmosphérique, nous avons conçu et déployé dans la ville de Lyon un réseau de capteurs de pollution économe en énergie. Sur la base des caractéristiques de notre système et des jeux de données de la pollution atmosphérique, nous avons évalué l'efficacité de nos approches de déploiement et d'ordonnancement. Nous présentons et discutons dans cette thèse les résultats d'évaluation de performances ainsi que des lignes directrices pour la conception de systèmes de suivi de la pollution de l'air. Parmi nos principales conclusions, nous soulignons le fait que la taille optimale du réseau de capteurs dépend du degré de variation des concentrations de pollution dans la région de déploiement.

MOTS-CLÉS : Réseaux de capteurs sans fil, déploiement, ordonnancement, pollution de l'air.

Laboratoire (s) de recherche : CITI
Directeur de thèse: Hervé Rivano

Composition du jury :

DIAS DE AMORIM, Marcelo	Directeur de recherche, CNRS	Rapporteur
DUDA, Andrzej	Professeur des universités, Grenoble INP-ENSIMAG	Rapporteur
CARNEIRO VIANA, Aline	Chargée de recherche HDR, INRIA	Examinatrice
GUERIN-LASSOUS, Isabelle	Professeur des universités, Univ. Lyon 1	Examinatrice
RIVANO, Hervé	Professeur des universités, INSA de Lyon	Directeur de thèse
BECHKIT, Walid	Maitre de conférences, INSA de Lyon	Co-directeur de thèse
CHAPPAZ, Claire	Atmo Auvergne-Rhône-Alpes	Invitée